

Baseband Digital Loops

Whitham D. Reeve
Reeve Engineers, Inc.

I. Applications and Networks
II. General Considerations for Baseband Digital Loops
III. Baseband Signaling
IV. Repeatered T1-Carrier
V. Subrate Digital Loops
VI. Switched-56
VII. ISDN Digital Line
VIII. High-Bit-Rate Digital Subscriber Line

GLOSSARY

Baseband Signals in their unmodulated form, usually including spectral components near zero frequency (dc).
Baud Signaling rate of a data signal in symbols per second.
Bit rate Information rate of a data signal, not necessarily equal to baud, measured in bits/second (b/s).
Crosstalk Unintentional transfer of a signal in one circuit to an adjacent circuit.
Digital Signaling with elements from a finite set.
Digital subscriber line Twisted-pair loop used for the transmission of digital signals.
Line coding Formating of information into digital symbols and assigning pulse waveforms to each symbol.
Loop Telecommunication facility between the user and the serving central office.
Power spectral density Distribution of power in the frequency domain, typically expressed in dBm per Hz of bandwidth (dBm/Hz).

THE 19TH-CENTURY telegraph was a form of baseband digital signaling using metallic media (open wire) as the transmission line. It involved the transfer of information by turning the line current on and off at different rates ("dots" and "dashes") to indicate different characters of a message, a form of on-off keying (OOK). The transmission rate of the early telegraph was low by today's standards, in the neighborhood of 10 b/s. As electronic telecommunication technologies evolved, baseband digital transmission became possible over significant distances on metallic coaxial media, reaching speeds of a few hundred megabytes per second, while baseband digital transmission over metallic twisted pairs reached speeds of a little over 3 Mb/s. Using optical fiber media, transmission speeds now reach 40 Gb/s, or 40 billion b/s, in the public network.

A baseband signal is unmodulated and has spectral components near zero frequency (dc). Voltage pulses applied directly to twisted pairs are a form of digital baseband signaling. Such a signal also may be used to modulate a carrier, in which case the original baseband signal is converted

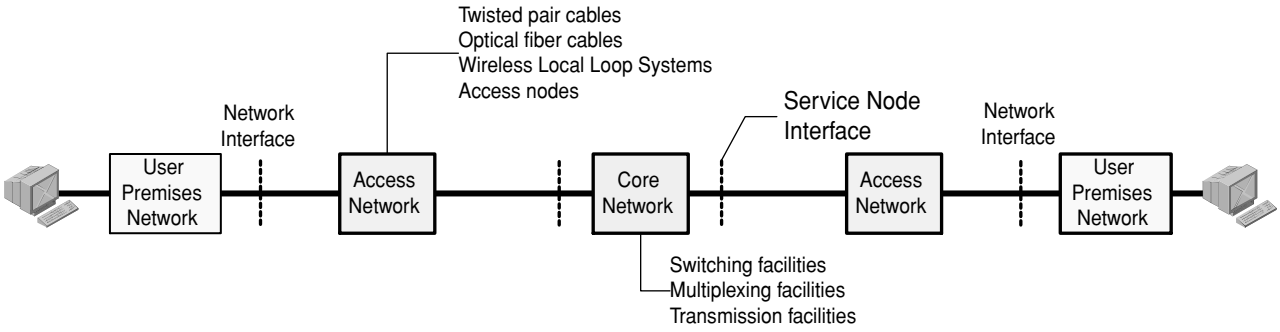


FIGURE 1 Services are transported from the core network to users via the access network. The access network itself does not provide telecommunication services.

to passband signal with the same information. One example of such a device that converts a baseband signal to a passband signal is the ubiquitous analog voiceband modems used to transfer a digital data stream through a telecommunication channel designed for a single analog voice channel.

A telecommunication channel consists of an access network at each end and a core network in between (Fig. 1). The access network is that part of the telecommunications network between the telecommunications subscriber's premises and the serving central office and is located at the edge of the core network. The access network traditionally has been called the subscriber loop or subscriber line. It primarily consists of metallic (copper) twisted-pair cables at this time but is evolving to a combination of twisted-pair cables, access nodes, and optical fiber cables. The metallic twisted pair is a very satisfactory transmission line and has well-understood and predictable characteristics. The access node collects traffic from individual subscribers and consolidates (by multiplexing) this traffic for transmission to the serving central office on a relatively high capacity transmission system (Fig. 2).

I. APPLICATIONS AND NETWORKS

Digital loops may be used in any application requiring the transfer of digitized information from one place to another as long as the constraints of the transmission medium are taken into account. The ubiquitous metallic twisted pair is most often used as the initial connection between the user and the public network. Higher capacity transmission systems are then used to carry the combined traffic of many users. Typical applications of digital loops are the interconnection of:

- Local area networks (LAN) in different geographical locations
- Automatic teller machines and a central banking center
- Airline ticket counter terminals and a central reservations center
- Point-of-sale terminals (cash registers and credit card authorization machines) and central sales locations or banks
- Air traffic control consoles and remote airport or air route traffic control radars
- Remote digital subscriber line access multiplexers (DSLAM) and Internet service provider routers
- Business establishments requiring voice and data services over a single loop

The bit streams from many different applications are consolidated by digital loop carrier (DLC) systems, access nodes, and primary multiplexers (channel banks). The outputs of these systems are further consolidated by a variety of metallic twisted-pair, microwave radio, and optical fiber transmission systems operating at data rates of 1.544 Mb/s (in the United States) and higher for transmission throughout the telecommunications network.

The telecommunication networks of industrialized countries are evolving toward all-digital, multimedia networks that provide end-to-end carriage of data, digitized voice, and digitized video traffic. Some services involve integrated handling of different media and extensive user signaling to and through the network so users have access to a variety of services. Such services are handled through an integrated services digital network (ISDN), which is a concept for a family of standard digital interfaces and services. The ISDN is not a new or separate network; it is evolving from today's existing analog and digital networks.

An emerging application for digital loops is the so-called voice over digital subscriber line (VoDSL). This type of DSL, more formally referred to a SHDSL (symmetric high-bit-rate DSL or, alternatively, single-pair high bit-rate DSL), provides rate-adaptive digital service over one or two twisted cable pairs. The entire transmission bandwidth is used for digital connectivity and is

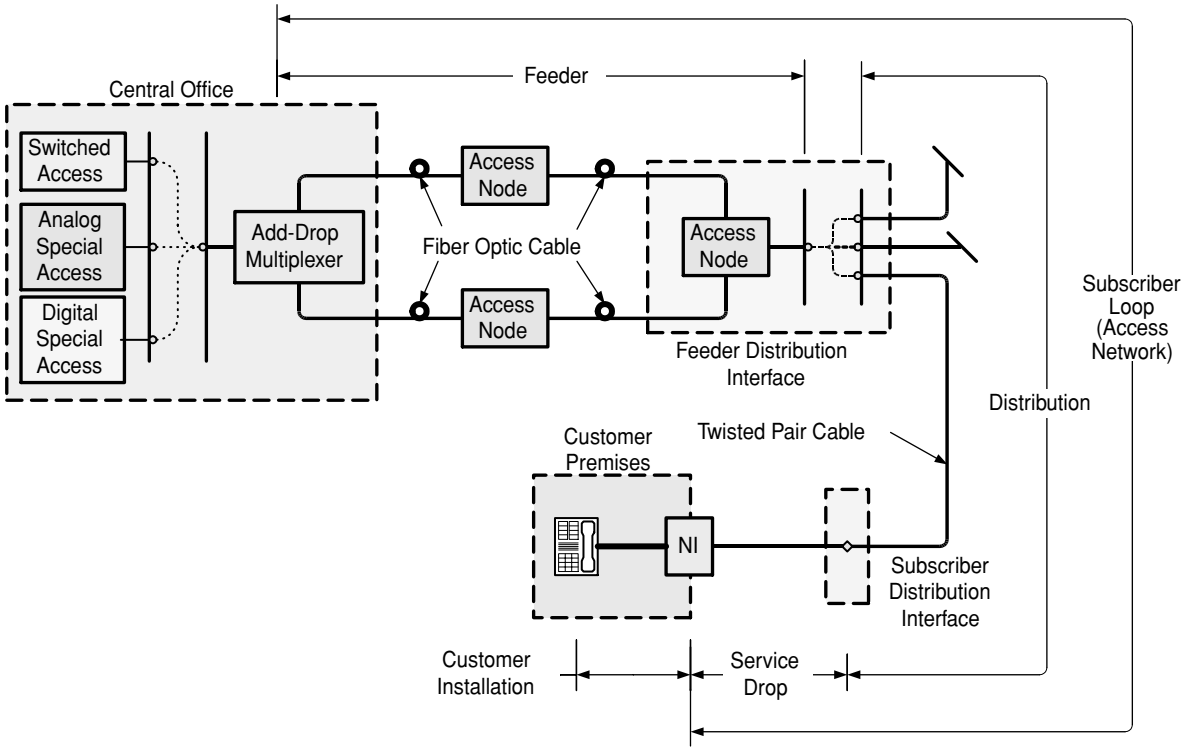


FIGURE 2 Access node application in the subscriber loop. Access nodes may be simple primary multiplexers (channel banks) or intelligent digital loop carrier remote terminals. Many different technologies, including optical fibers (shown), twisted pairs, and digital wireless systems, may be used to connect access nodes to the central office terminal, but twisted pairs are the most common from the access node to the user.

dynamically allocated to multiple voice and data channel applications as required (Fig. 3). SHDSL, which uses baseband transmission techniques, differs from the asymmetric DSL (ADSL or, more commonly, xDSL). ADSL uses passband transmission techniques and provides a high-speed digital connection plus regular analog telephone service (plain old telephone service, or POTS) on a single cable pair (commonly called "line sharing"). With ADSL, analog bandwidth from dc (zero frequency) up to around 4 kHz is used for POTS while higher frequencies are used for the bidirectional digital service. A splitter, or filter, is required to separate the two frequency bands and allow simultaneous operation of analog and passband digital service over the same twisted pair.

II. GENERAL CONSIDERATIONS FOR BASEBAND DIGITAL LOOPS

A baseband digital loop is a twisted-pair subscriber loop that transmits voltage pulses between the serving central office or access node and the user. The pulses represent the values of bits or groups of bits and are shaped for optimal transmission on twisted pairs. There are many types of twisted-pair digital loops in use, including:

- ISDN digital line, both basic rate interface (BRI) and primary rate interface (PRI)
- Subrate digital loops (both switched and dedicated)
- Repeatered T1-carrier
- Many variations of the digital subscriber line (xDSL), including the high bit-rate digital subscriber line (HDSL), single-pair HDSL (SHDSL), and asymmetric DSL (ADSL)

Of the many xDSL variants, only the HDSL and its derivatives (HDSL2 and SHDSL) use baseband technology. All other types of xDSL use passband techniques and are not discussed here except to put them in context with other digital loops. Note that the "x" in xDSL can be almost any letter limited only by the ingenuity of equipment manufacturer's marketing departments.

All digital loops can potentially interfere with other digital loops in the same cable through cross-talk coupling. Therefore, the development of any new loop technology requires careful coordination with current and older technologies through a process of spectrum management and the service and technology administration of metallic twisted pairs in the same cable. Generally, spectrum compatibility is assured through the design of receivers

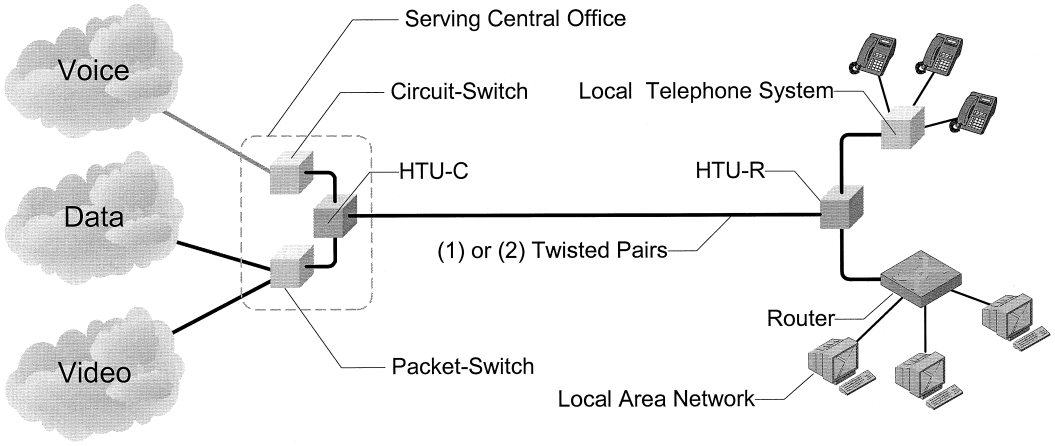


FIGURE 3 Symmetric HDSL application allows multiple voice and data connections by dynamically allocating bandwidth. While this type of application is possible with other DSLs described in this article, SHDSL was originally designed for it.

that are insensitive to interference, through the design of transmitters with signal power limits and spectral shapes that do not degrade other loops, and through technology deployment guidelines. The cross-talk characteristics of twisted pairs are considered invariable design parameters because of the huge embedded base of installed plant. Among the deployment guidelines are recommendations to avoid mixing certain loop types in the same binder group in a cable.

The digital loops commonly used today are listed in Table I along with the line and payload rates. Payload rates are the digital transmission rates available to the user, while line rates include overhead for network control and framing. Each loop type will be described in greater detail later. Almost all baseband digital loops were conceived long before the International Standards Organization (ISO) developed the protocol layer concept defined by the open systems interconnect (OSI) model; however, for purposes of comparison, the descriptions in this article can be considered applicable to the OSI Layer 1, or physical layer.

III. BASEBAND SIGNALING

In general, baseband signals on loops are restricted to balanced bipolar pulses. The spectral characteristics of baseband signals are fundamentally defined by their shape, including pulse interval, amplitude, and width. Since rectangular voltage pulses have a wide spectrum, the requirements for all digital loops specify filters to control the spectrum. The process of tailoring signals for transmission on a particular media and associating bit values with specific signal voltages is called *line coding*. Some of the requirements for line coding on twisted-pair exchange cables include:

- Elimination of zero frequency (direct current, dc) and very low frequencies from the signal spectrum: This allows the use of coupling transformers, which do not pass dc or very low frequencies and which isolate the electronic interfaces from the hostile outside plant cable environment. It also allows the use of dc for line powering and simple network control functions.
- Self-clocking of the bit streams: The data receiver can recover the clock from the incoming signal and then determine where to read the bit or symbol value of the signal. The line code must provide sufficient pulses not only on an average basis but over some specific time interval as well.
- Error detection: It is necessary to determine the performance of digital loops at the physical layer for testing and monitoring purposes. Some line codes provide a means for detecting errors (but not correcting them) without introducing additional error-detection bits into the data stream. Other line codes include error detecting and forward error correcting methods.
- Bandwidth minimization: Multilevel line codes increase the bandwidth efficiency by reducing the required bandwidth for a given data rate. In digital loop applications, the bandwidth reduction is manifested as a reduction in signal energy at higher frequencies, where twisted pair attenuation is higher. With multilevel line codes, the amplitude of a pulse represents the value of a bit or group of bits. Another benefit of bandwidth minimization is in the reduction of cross-talk interference in which

TABLE I Digital Loops

Digital loop	Payload rate	Line rate	No. of cable pairs	Type
Dedicated Digital Service subrate digital loop	2.4 to 64 kb/s	2.4 to 72 kb/s	2	Baseband
Switched 56 (type I)	56 kb/s	56 kb/s	2	Baseband
Switched 56 (type III)	56 kb/s	160 kb/s	1	Baseband
ISDN digital subscriber line (BRI)	144 kb/s	160 kb/s	1	Baseband
High-bit-rate digital subscriber line (HDSL)	1.536 Mb/s	Two loops each operating at 784 kb/s	2	Baseband
High-bit-rate digital subscriber line (HDSL2)	1.536 Mb/s	One loop operating at 1.552 Mb/s	1	Baseband
Symmetric rate-adaptive HDSL (SHDSL)	192 kb/s to 2.304 Mb/s or 384 kb/s to 4.608 Mb/s	192 kb/s to 2.312 Mb/s or 384 kb/s to 4.624 Mb/s	1 2	Baseband Baseband
Repeatered T1-carrier	1.536 M/bs	1.544 Mb/s	2	Baseband
Asymmetric digital subscriber line (xDSL)	32 kb/s to 6.4 Mb/s downstream; 32 to 640 kb/s upstream	Up to approximately 8 Mb/s	1	Passband
Very-high-speed digital subscriber line (VDSL)	5.792 to 43.336 Mb/s downstream; 1.448 Mb/s to 5.792 Mb/s upstream	6.480 to 51.840 Mb/s downstream; 1.620 to 6.480 Mb/s upstream	1	Passband

the signals on one cable pair interfere with the signals on another. Cross-talk coupling in twisted-pair exchange cables is higher at higher frequencies, so reducing the energy at the higher frequencies also reduces cross-talk interference.

- Power spectral density (PSD) control: Control of transmitted signal power spectral density is required to reduce cross-talk interference.
- Noise immunity: Twisted-pair exchange cables exist in a noisy environment. Some line codes have higher noise immunity than others, and the ones with the higher immunity are chosen specifically for loop applications.

Table II lists the line codes used with each digital loop. Most are a form of the fundamental return-to-zero (RZ) line code. This list is not exhaustive but represents the line codes used with loops discussed in this article. Some loop types modify a basic line code for specific purposes such as network control or to overcome inherent clock recovery limitations. Rather than provide a general discussion of how each line code works here, each loop type will be described along with the specific line code and any particular modifications.

Baseband signals also include those used at the interfaces of data terminal equipment (DTE) with data communication equipment (DCE). Typically, unipolar or bipolar signals are used to convey data values and control signals over the short distances between DTE and the DCE. Generally, these types of interface signals are not suitable for direct transmission over twisted-pair exchange cables because of electrical protection requirements, signal incompatibility, and interference (cross-talk) problems. Data service units (DSU) and channel service units (CSU) convert the DTE signals to a more suitable type of baseband signaling.

The bandwidth required by baseband signaling is directly related to the data speed by the Nyquist bandwidth constraint: A system with bandwidth $W = 1/2T$ Hz can support, without intersymbol interference, a maximum transmission rate of:

$$R_s = 2W \tag{1}$$

where R_s is the transmission rate in symbols per second and T is the symbol interval in seconds.

It is necessary that the baseband digital loops not be connected to any circuitry that limits the bandwidth to

TABLE II Baseband Digital Loop Line Codes

Digital loop	Line code
Repeatered T1-carrier	Alternate mark inversion (AMI) or bipolar eight zero substitution (B8ZS)
Subrate digital loop (SRDL)	Alternate mark inversion (AMI) with modifications
Switched 56 (SW56)	Alternate mark inversion (AMI) with modifications
ISDN digital subscriber line (ISDN DSL)	2-binary, 1-quaternary (2B1Q)
High-bit-rate digital subscriber line (HDSL)	2-binary, 1-quaternary (2B1Q)
HDSL2 and SHDSL	Trellis-coded 16-level pulse amplitude modulation (TC-PAM)

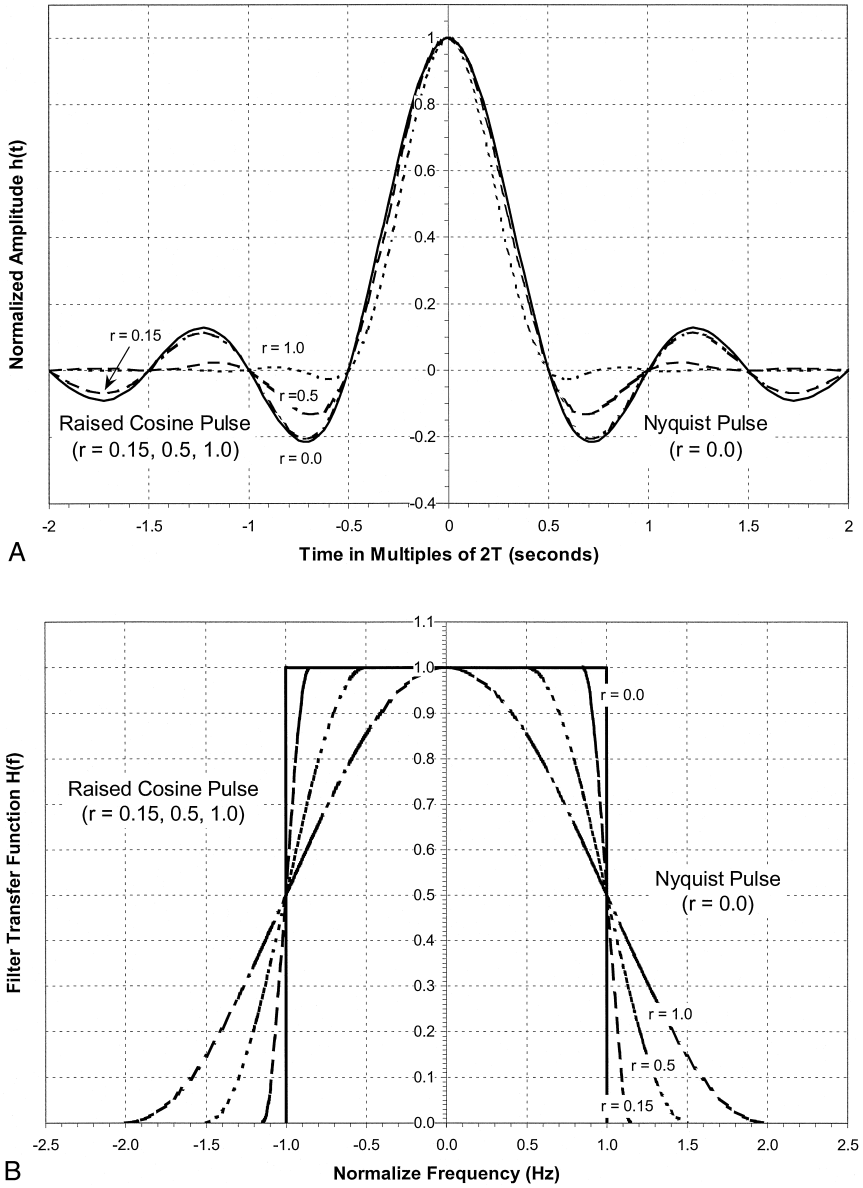


FIGURE 4 Band-limited pulses that do not contribute intersymbol interference. (a) Nyquist pulse and raised cosine pulses; (b) frequency characteristics of pulses shown in (a); (c) consecutive raised cosine pulses do not cause intersymbol interference because they have zero amplitude at the sampling instance.

something less than the specified value. Also, most baseband digital loops require twisted pairs that are dc continuous for line powering or loopback control. None of the digital loops discussed in this article is compatible with analog voiceband channels because of the latter's limited bandwidth and intervening electronics that preclude dc continuity.

A rectangular pulse wholly contained within one symbol interval of T seconds does not produce intersymbol interference (ISI). However, such a pulse requires more bandwidth than is necessary and requires very accurate sampling of the received signal exactly at pulse peaks. The Nyquist pulse (Fig. 4a) satisfies the ISI condition with the minimum possible bandwidth of $1/2T$ Hz (a normalized frequency of 1.0 in the filter transfer function response of Fig. 4b). Since it is difficult to generate and work with pulses that have such an abrupt frequency spectrum (roll-off factor $r = 0.0$), actual interfaces will use a pulse with slightly greater bandwidth and smooth frequency characteristics while retaining the property of zero samples from

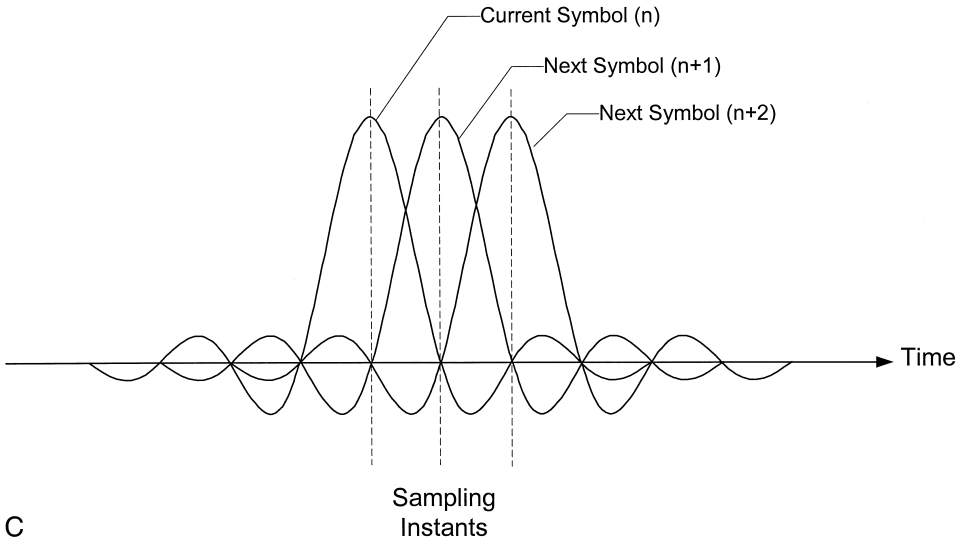


FIGURE 4 (*Continued*)

other pulses at symbol intervals. The frequency characteristic of a raised cosine pulse (Fig. 4a) is often used. The edge of the frequency characteristic is a cosine function rolling off in an additional fraction of the minimum $1/2T$ Hz bandwidth. (In addition to the Nyquist pulse with $r = 0.0$, roll-off factors of 0.15, 0.5, and 1.0 are shown in Fig. 4.) The selection of a filter (and, thus, the pulse shape) depends on the relative noise and bandwidth constraints in a given application. Older digital loops used relatively wide filters with Butterworth characteristics, while newer digital loops use programmable digital filters.

The Nyquist bandwidth constraint of Eq. (1) can be modified to account for the filter roll-off factor:

$$R_s = 2W/(1 + r) \tag{2}$$

where R_s and T already have been defined, and r is the filter roll-off factor.

An example will illustrate the application of Eq. (2). The ISDN basic rate interface is a baseband digital loop that uses a four-level pulse amplitude modulated (PAM) line code operating at a rate, R_d, of 160 kb/s and having a transfer characteristic with 25% excess bandwidth. Since the number of levels $M = 2^k = 4$, then each symbol consists of two bits and $k = 2$. Therefore, $R_s = R_d/k =$ 80 kilosymbols/s = 80 kbaud. The minimum required bandwidth from Eq. (2) is $W = R_s(1 + r)/2 = 50$ kHz.

The performance of all twisted-pair digital loops, in terms of bit error rate (BER) is identical. Loops designed and maintained according to telecommunications industry practices will provide a BER of 1×10^{-7} (one received bit error in 10 million transmitted bits) with a 6-dB margin. During normal operation, when a 6-dB margin is attained, the BER actually will be 4 to 6 orders of magnitude better. However, the margin takes into account unpredictable interference from other loops sharing the same cable (crosstalk) as well as impulse noise and noise from other sources. It is under these conditions that the BER will be (at least statistically) no worse than 1×10^{-7}.

IV. REPEATERED T1-CARRIER

Repeatered T1-Carrier was developed in the early 1960s for interoffice applications but soon found its way into the loop. Repeatered T1-carrier operates at the DS-1 rate (1.544 Mb/s) and uses digital baseband transmission over two cable pairs, one for transmit and one for receive. With proper loop design, this technology can be used on loops as long as 200 miles, although the typical loop is much shorter. Repeatered T1-carrier is a very reliable technology but is characterized by extensive engineering and installation effort when done properly. There is a huge embedded base of T1-carrier in the public network and its use will continue for years to come.

The main components of repeatered T1-carrier are central office repeaters, which are installed in a line termination shelf in the central office or DLC remote terminal, and line repeaters, which are installed in repeater housings in the outside plant (Fig. 5). Housings are available to hold from one to as many as 50 line repeaters and are spaced approximately 5000 to 6000 ft apart. End sections are about 2/3 the length of repeatered sections to account for higher impulse noise in cables near the central office or

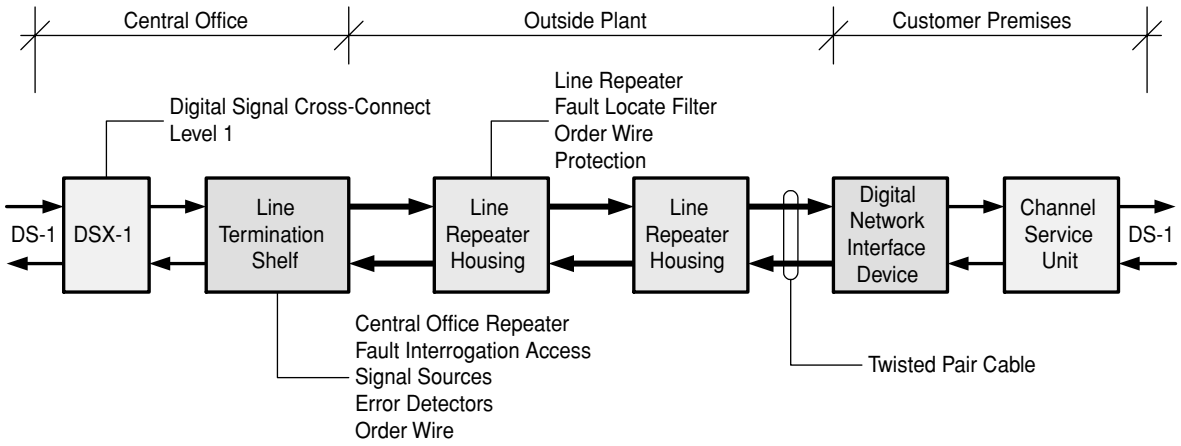


FIGURE 5 Repeatered T1-carrier components. Two cable pairs are required for transmission, one for each transmission direction.

subscriber premises. Other ancillary components, such as fault interrogation units and fault locate filters, are used for troubleshooting. When the carrier system terminates on a user's premises, the network operator normally will install a digital network interface device, which functions like a central office repeater. A channel service unit owned by the subscriber provides the interface between user equipment and the span line. Repeatered T1-carrier can be used in any application requiring the transport of a DS-1 rate signal. Figure 6 shows an application where repeatered T1-carrier is used to connect a customer's private branch exchange (PBX) with a central office switching system.

A DS-1 rate signal consists of 24 DS-0 rate channels. Each DS-0 channel has eight bits, so each DS-1 data frame has 192 bits. Each DS-0 channel operates at 64 kb/s giving a total payload rate of 24×64 kb/s, or 1.536 Mb/s. A framing bit precedes the 192 data bits. In its simplest form, with the superframe (SF) frame structure, the framing bits not only set frame boundaries but also identify optional embedded signaling bits (A-bit and B-bit signaling) associated with each DS-0 channel. The extended superframe (ESF) frame structure provides additional capabilities, including an embedded operations channel, cyclic redundancy check (CRC), and two additional signaling bits (C-bit and D-bit) associated with each channel. A technique called *robbed bit signaling* uses the optional A-, B-, C-, and D-signaling bits.

Repeatered T1-carrier uses the alternate mark inversion (AMI) line code. The AMI line code ensures that the signal carries no undesired dc component, as dc is used for line powering. With AMI, a binary-zero is transmitted as no pulse, or zero volts. A binary-one is transmitted either as a 50% duty cycle positive or negative pulse, opposite in polarity to the previous pulse (Fig. 7). The pulse width, pulse period, and pulse amplitude are given in Table III. Before the pulses are actually coupled to the twisted pair loop, they are filtered in the transmitter line driver with a low-order Butterworth filter that rounds off the pulse edges and may introduce some overshoot on the leading and trailing edges.

The AMI line code has no inherent capability to ensure adequate pulse density for clock recovery. Clock recovery for repeatered T1-carrier requires an average pulse density of 12.5% and no more than 15 consecutive pulse periods with no pulses. User payload, including digitized speech, can have long strings of binary-zeros, which will violate

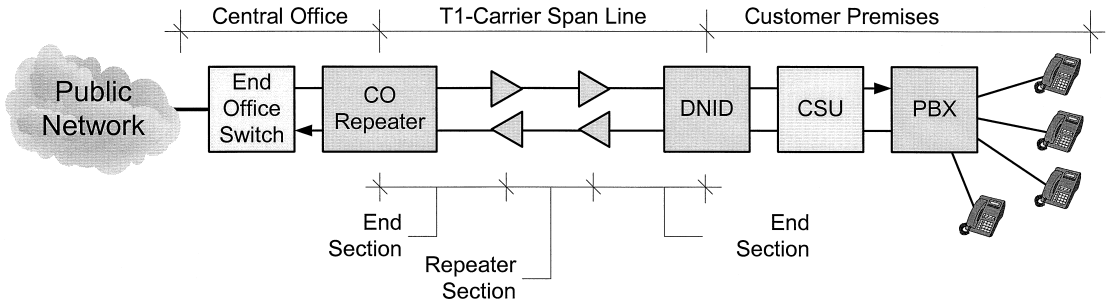


FIGURE 6 Typical repeatered T1-carrier application. A DS-1 circuit has 24 channels and provides up to 24 trunks from the private branch exchange to the public network.

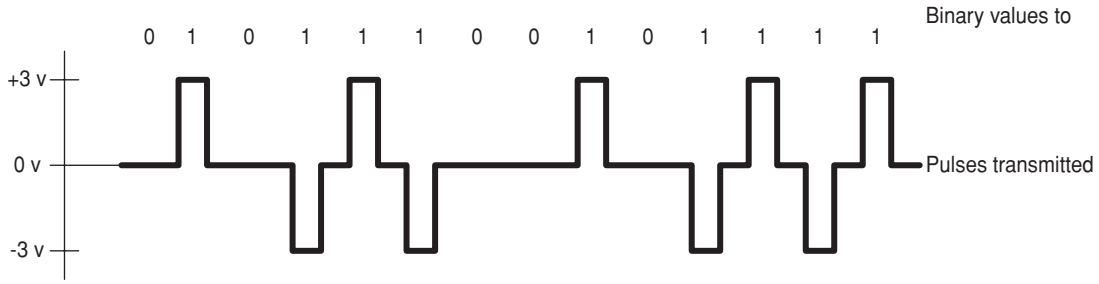


FIGURE 7 Alternate mark inversion (AMI) line code. Voltage pulses representing binary-one values alternate in polarity. Binary-zero values are transmitted as zero voltage.

these requirements. For speech, the payload is modified by substituting a binary-one in one of the bit positions, which increases the pulse density at the output of the line transmitter. Occasionally changing bit values has no perceptible effect on speech but clearly is unacceptable for data, so a different method must be used. For data, the number of user bits is reduced to seven (from eight), which restricts the data payload rate to 56 kb/s when AMI is used. The eighth bit is set to a binary-one, which guarantees the pulse density constraints will be met.

Where clear channel capability (CCC) is required, in which each channel supports eight user bits and a full 64 kb/s payload rate, the AMI line code is modified to provide intentional bipolar violations (BPV). This ensures adequate pulses for clock recovery. For repeatered T1-carrier, this modification is called B8ZS, or bipolar with eight zero substitution. With B8ZS, four pulses—two intentional BPVs and two normal pulses—are substituted for any string of eight binary-zeros (**0000 0000**, or no pulses). Intentional BPVs are inserted in bit positions four and seven. Each BPV is followed by valid pulses (to ensure there is no residual dc component) in the pattern **000V B0VB**, where

0 = Zero volts transmitted, binary-zero (original user data).
V = ±3.0 v–p with polarity in violation of the AMI line code, binary-one.
B = ±3.0 v–p with polarity determined by the normal AMI line code, binary-one.

The known sequence **000V B0VB** is recognized in the line receiver and replaced with a string of eight binary-zeros at the user interface (**0000 0000**). Figure 8 shows the AMI with B8ZS line code.

TABLE III T1-Carrier Pulse Parameters

Parameter	1.544 Mb/s
Pulse period	648 ns
Pulse width	324 ns
Pulse amplitude	±3.0 volts–peak (v–p)

To ensure backward compatibility of current production equipment using modern technology with the huge embedded base of older equipment, the T1-carrier span line design requirements have remained unchanged since the 1960s. The design process is rather involved and requires close scrutiny of many variables, including cable types, installation conditions, and the effects of cross-talk.

T1-carrier was designed as a repeatered system from the outset. Each span line section terminates in a line repeater (Fig. 5). A line repeater consists of two signal regenerators, one for each transmission direction. A regenerator detects the incoming signal and decides if a pulse or no pulse is present at sample time. If a pulse is present, the regenerator reconstructs a new pulse of the same polarity as the incoming pulse and transmits it. If no pulse is present, the regenerator transmits zero voltage. As many as 200 repeatered sections are possible (limited by the accumulation of jitter), giving a maximum T1-carrier span line length of approximately 200 miles.

V. SUBRATE DIGITAL LOOPS

Subrate digital loops operate at standard speeds from 2.4 to 64 kb/s with the higher speeds of 56 and 64 kb/s being the most common. These loops were first deployed in the mid-1970s and require two metallic twisted cable pairs, one for each direction of transmission (Fig. 9). The subrate digital loop originally was designed as the access portion of a dedicated, end-to-end, synchronous digital service variously called Digital Data Service, Dataphone® digital service, or dedicated digital service, all of which use the acronym DDS. DDS channels can be operated in the point-to-point and point-to-multipoint modes, except that the 64-kb/s user rate is point-to-point only. Most current applications are point-to-point only and frequently are used to access packet-switched network technologies such as frame relay. Even though the subrate digital loop is

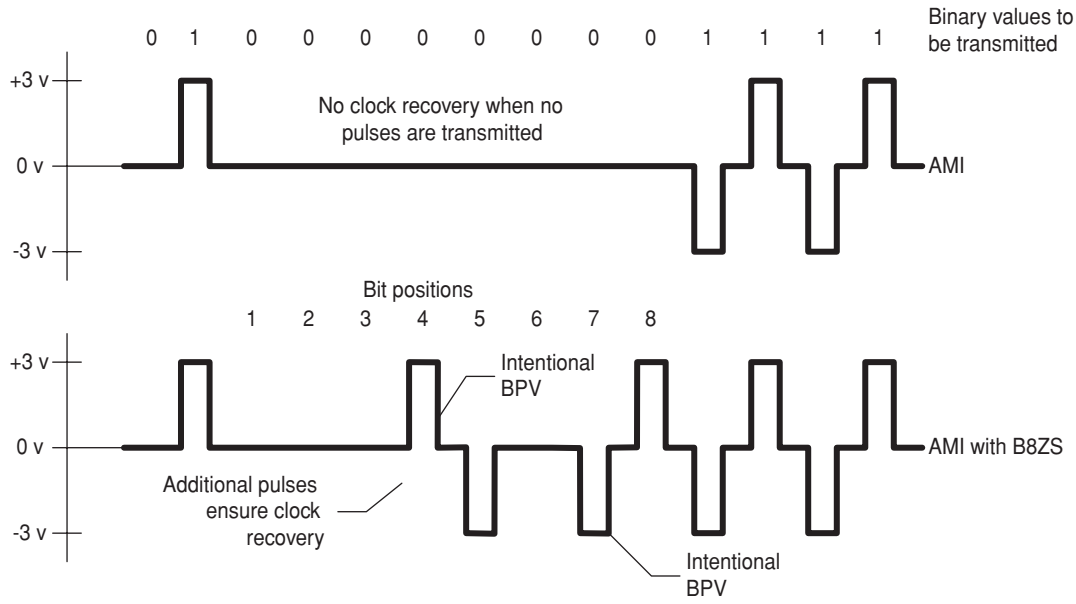


FIGURE 8 AMI with B8ZS uses intentional bipolar violations to provide the necessary pulse density for clock recovery when user data has long strings of binary-zeros. The upper trace shows how the example binary values would be transmitted with AMI, the lower trace shows the B8ZS for this binary sequence.

a fully synchronous loop, any standard asynchronous user rate up to 57.6 kb/s, usually in increments of 9.6 kb/s, can be accommodated by modern terminal equipment using standard rate adaptation methods.

The subrate digital loop normally terminates in a user's data service unit (DSU) at one end and in an office channel unit data port (OCU-DP) at the serving central office or access node. Normal range is 18,000 to 25,000 ft at 56 kb/s depending on cable gauge. An access node, which can be a simple channel bank installed in a remote cabinet, and regular digital transmission facilities may be used to extend the subrate digital loop beyond the normal transmission limits of the DSU and OCU-DP. An OCU-DP is not always required. Two cable pairs meeting the attenuation constraints can be used to connect two DSUs as shown in Fig. 9b.

All OCU-DPs provide up to 20 mA of sealing current on the subrate digital loop. This sealing current, in addition to keeping high-resistance oxide film from building up on cable splices, also provides a signaling path for the channel loopback (loopbacks are discussed later). Some OCU-DPs also are capable of providing dc powering current to

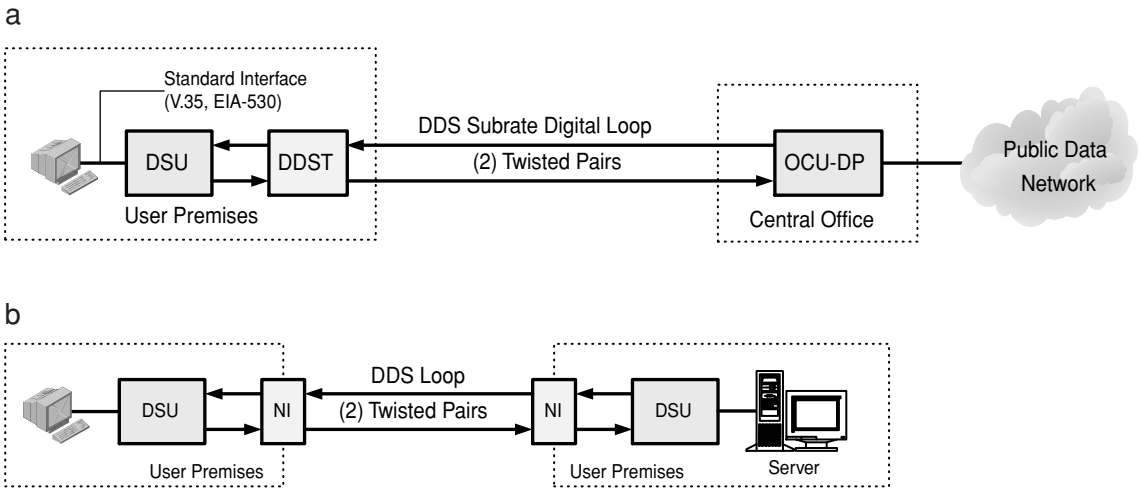


FIGURE 9 Dedicated digital service subrate digital loop. (a) The DDS loop allows access to the digital data network. (b) A simple local DDS loop requires only two outside plant cable pairs and a DSU at each end.

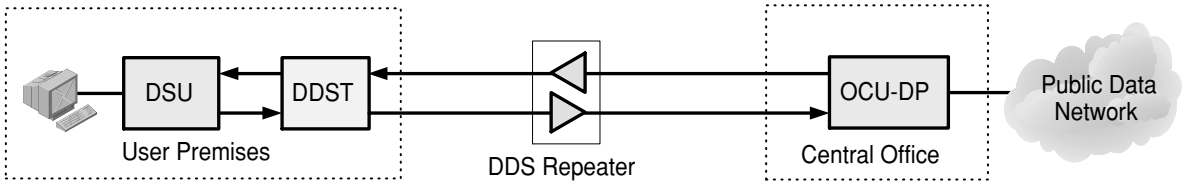


FIGURE 10 Subrate digital loop length may be extended with a DDS repeater. A digital data station termination (DDST) provides network loopback for troubleshooting the access network connection.

repeaters and digital data station terminations (DDSTs). The repeater regenerates the signals and effectively doubles the loop length between the central office and the user (Fig. 10).

The block diagram in Fig. 11 shows the main functional components in a DSU. The DSU serves two distinct purposes. The first purpose is to isolate the customer installation (CI) from the network, thus protecting the network from possible harmful problems in the CI. The second purpose of the DSU is to provide standard interfaces to user DTE. Interface functions are shown on the left side of Fig. 11. In particular, most DSUs provide EIA-232, EIA-530, and V.35 interfaces with one or more standard rates. The DSU adapts the user rate to the line rate. For example, a synchronous line rate of 56 kb/s could be used with a user rate of 19.2 kb/s (either synchronous or asynchronous) through rate adaptation.

Table IV lists the standard DDS user rates and the actual line rates on subrate digital loops without a secondary channel (a secondary channel provides a low-speed signaling subchannel, which is shared between the user and network, the latter for control purposes). This table also shows the Nyquist frequency, which is one-half the line rate and the frequency used for loop design. The user rates and line rates are the same for the speeds through 56 kb/s if there is no secondary channel involved. When a secondary channel is used, the line rate increases slightly to include the user rate plus additional overhead for the shared network control and secondary channel bits. For the 64 kb/s user rate, the line rate is always 72 kb/s, which includes the user rate and 8 kb/s for framing. No secondary channel is available with the 64 kb/s user rate.

Data bits are carried on the 56-kb/s subrate digital loop in an unstructured format, with seven data bits, D1, D2, D3, D4, D5, D6, and D7. The least significant bit (D1) is sent first. In the direction from the subrate digital loop to the network, the OCU-DP adds a binary-one to the seven data bits and thus provides an 8-bit word to the network DS-0 channel. In the other direction, the OCU-DP removes this eighth bit and sends only seven bits out onto the subrate digital loop.

The 64-kb/s subrate digital loop uses 9 bits with the loop-byte structure of D1, D2, D3, D4, D5, D6, D7, F, D8. The F-bit is a framing bit which follows the repeating

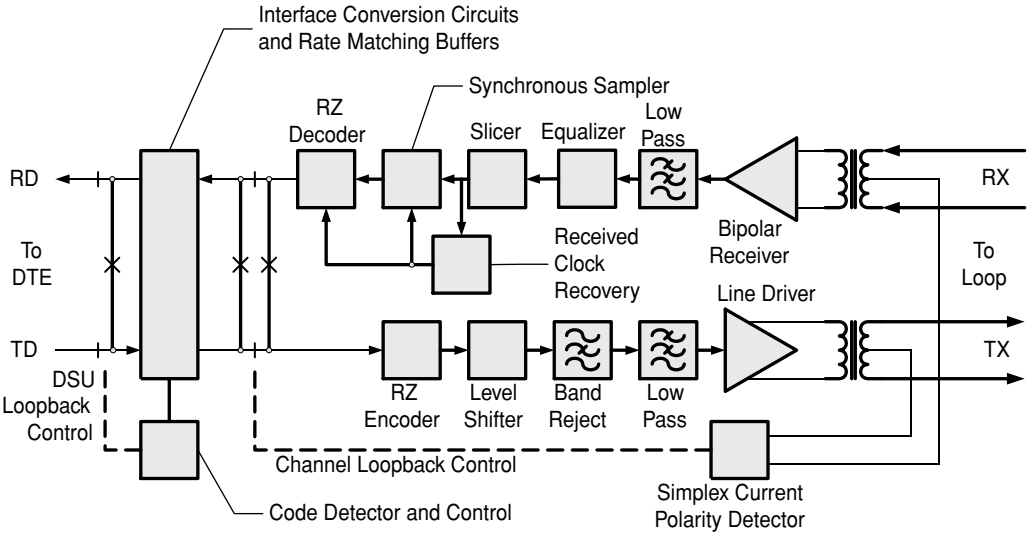


FIGURE 11 Data service unit block diagram. The block diagram for the OCU-DP at the other end of the loop is similar except it does not have interface conversion circuits. Instead, the OCU-DP couples the recovered data to the PCM bus in a multiplexer.

TABLE IV Standard Subrate Digital Loop Rates

User rate (kb/s)	Line rate (kb/s)	Nyquist frequency (kHz)
2.4	2.4	1.2
4.8	4.8	2.4
9.6	9.6	4.8
19.2	19.2	9.6
38.4	38.4	19.2
56.0	56.0	28.0
64.0	72.0	36.0

pattern 101100 . . . . The OCU-DP strips this bit before the data are formed into the DS-0 channel. Similarly, the DSU strips the F-bit from the received data before the data are converted to conform to the DTE interface requirements.

The subrate digital loop uses the AMI line code, which is modified under certain circumstances. The pulse width, pulse period, and pulse amplitude are given in Table V. Two consecutive pulses with the same polarity indicate a line code violation (BPV). However, intentional BPVs are used to convey network control information and to suppress strings of more than six or seven consecutive zeros. The following notations apply to the subrate digital loop:

0 = Zero volts transmitted, binary-zero.
B = ±1.40 v–p with polarity determined by the normal AMI line code, binary-one.
V = ±1.40 v–p with polarity in violation of the AMI line code, binary-one.
X = Either **0** or **B** depending on the need for the desired polarity of a BPV.

Any intentional line code violations must also ensure that there is no latent dc component; therefore, in the line code used with the subrate digital loop, a time slot is reserved (designated by the symbol **X**) prior to any intentional violations. This time slot is for application of a pulse or no pulse in such a way that successive violations (designated by the symbol **V**) alternate in polarity. The desired polarity of intentional violations (**V**) is achieved by assigning a value **0** or **B** to **X** such that the total number of **B**s since the last **V** is odd. Because it is not desirable to have adjacent pulses of the same polarity, the **X** and **V** time slots are separated by a no-pulse condition, which gives an easily recognizable **X0V** pattern in each intentional BPV sequence. Table VI shows several control code sequences that use the **X0V** pattern (this table does not show all possible code mapping sequences). These patterns are repeated as long as the control code condition is being transmitted.

TABLE V Pulse Parameters

Parameter	56 kb/s	64 kb/s
Pulse period	17.86 μs	13.89 μs
Pulse width	8.93 μs	6.94 μs
Pulse amplitude	±1.40 v–p	±1.40 v–p

TABLE VI Network Control Code Sequences for 56-kb/s and 64-kb/s Rates

Sequence	Pattern
User-to-NI direction	
Idle code	BBBBX0V
Zero suppression	0000X0V (substituted for 0000000)
NI-to-user direction	
Idle code	BBBBX0V
Zero suppression	000X0V (decoded as 000000)
Out-of-service	S00BX0V
Out-of-frame	S0BBX0V
DSU loopback	00B0X0V

Note: S indicates a "don't care" bit value.

Loopbacks provide for remote testing and are helpful in isolating problems between the network and customer installation. A loopback returns the signal to the sender (Fig. 12). When loopbacks are initiated from within the network (at the DS-0, or 64-kb/s, level), particular bit patterns, or network codes, are used. Where appropriate, these bit patterns are converted to local loop codes by the OCU-DP. The local loop codes are then sent on the subrate digital loop to the DSU. Similarly, when loopbacks are initiated at the DSU, particular local loop codes are sent on the subrate digital loop to the OCU-DP. The OCU-DP then converts these local loop codes to network codes for transmission within the network.

VI. SWITCHED-56

Switched-56 is an end-to-end digital, 56-kb/s, full-duplex, synchronous, circuit-switched service. Transmission is entirely over digital facilities; no intervening analog facilities may be used. Switched-56 service was originally deployed in the 1980s as a locally switched version of the 56-kb/s subrate digital loops previously described. Two additional versions were developed later that provided the same user rate but used different line codes.

Switched 56 service is more formally known as public switched digital service (PSDS), but in this section it will be called SW56 for clarity. SW56 includes three types of interfaces, type I, II, and III, which are summarized in Table VII. Types I and III are the most common.

Type I SW56 has the same working range as DDS subrate digital loops. Types II and III SW56 have shorter range

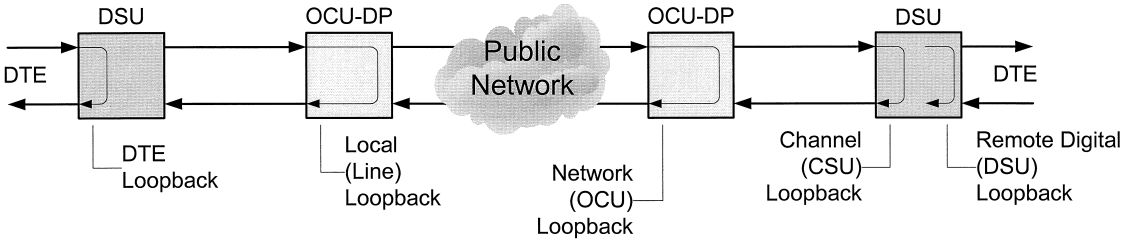


FIGURE 12 Dedicated digital service loopbacks. Alternate names for the various loopbacks are shown in parentheses.

but are usable to 18,000 ft on 24-ga. cable (14,000 ft on 26-ga. cable). If necessary, regular digital carrier facilities may be used to extend the SW56 beyond the transmission limits of the central office SW56 line circuit. As with the dedicated subrate digital loop, any asynchronous rate in multiples of 9.6 kb/s up to 57.6 kb/s and any standard synchronous rate up to 56 kb/s can be accommodated. Figure 13 shows a typical SW56 implementation.

The DSU used with Switched 56 service is called a switched-circuit DSU (SCDSU) to differentiate it from a DSU used for dedicated digital services. The SCDSU is the interface between the DTE and the loop and provides signal format conversion and rate adaptation. On the network side, the SCDSU will have a 2-wire or 4-wire interface depending on the SW56 type. On the DTE side, the SCDSU will have an EIA-232, EIA-530, or V.35 interface for data communication. The OCU-DP used with SW56 is operationally equivalent to the OCU-DP used with DDS.

Industry standards provide for interworking between the three SW56 interfaces, such that any interface type may call any other type and pass data under most conditions. When a call is established, a handshaking protocol takes place to determine whether a compatible protocol is available at the far-end SCDSU. If a compatible protocol exists, parameters are exchanged to indicate whether the data are to be sent synchronously or asynchronously and the data rate to be used. If the exchanged parameters are incompatible, then a conflict exists and data cannot be transferred. Interworking is also possible between SW56, the ISDN, and analog voiceband modems; however, interworking of different services involving subrate multiplexing, rate adaptation, packet switching, or the ISDN depends on the end-office switching system feature set and the capabilities of the end-user's SCDSU.

TABLE VII SW56 Interfaces

Type	Metallic interface	Signaling	Line rate (kb/s)	Payload rate (kb/s)
I	4-wire	In-band	56	56
II	2-wire	In-band	144	56
III	2-wire	Out-of-band	160	56

All SW56 interface types provide full-duplex transmission, but the technologies are very different. Type I SW56 uses a 4-wire metallic interface with a separate transmit and receive path, each operating at the 56-kb/s line rate. The line rate equals the payload rate because no out-of-band overhead is required for signaling. Signaling is via in-band bipolar patterns that include intentional bipolar violations. Type I uses a modified AMI line code with intentional bipolar violations for zero suppression and network control. The local loop signals are equivalent to the 56-kb/s DDS without secondary channel as previously described. However, the network path signals are different than DDS because they need to accommodate the in-band supervision and address signaling associated with a type I SW56 connection.

Type II SW56 uses a 2-wire metallic interface that operates in an analog mode and a digital mode. The analog mode is used for supervisory and address signaling. Once an end-to-end connection is established using the analog mode, the SCDSU is switched to the digital mode. Type II uses time compression multiplexing (TCM), also called "ping-pong." Bursts of data are sent over the local loop from one end at a 144-kb/s rate using the AMI line code. There is a dead time between bursts so the other end of the local loop can send a burst of data at 144 kb/s rate. Thus, each end of the local loop alternates between a transmit and receive state to achieve full-duplex transmission. The data sent over the network between line circuits are smoothed in buffers at a constant 56-kb/s rate.

Type III SW56 also uses TCM and the AMI line code on a 2-wire metallic interface; however, it only operates in the digital mode and the burst rate is 160 kb/s. Imbedded in each burst are two channels, one for signaling and the other for data. The signaling channel is for supervision and address signaling and operates at a rate of 8 kb/s. It is called an *out-of-band channel* because the bit positions for the signaling channel are separate from the bit positions for the data channel. The data channel operates at 64 kb/s. Each burst consists of a start and stop bit, 8 signaling bits, and 64 data bits, for a total of 74 bits. Table VIII

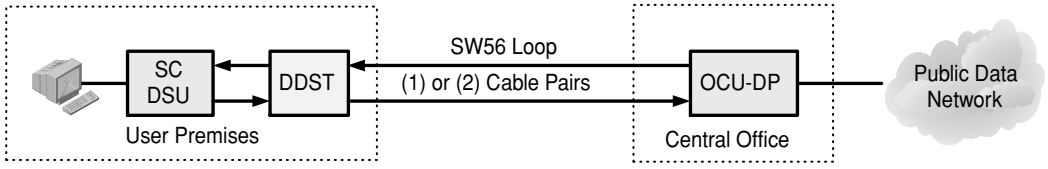


FIGURE 13 SW56 implementations are very similar to dedicated digital service subrate digital loops and can be used for the same purpose. Dialup typically is done at the SCDSU.

summarizes the electrical characteristics of each interface type.

A. SCDSU Originated Call

• *Idle*—Under idle conditions, the network provides a logic-0 to the SW56 line circuit or channel unit on the A-bit signaling channel embedded in a primary rate (DS-1) multiplexer. The line circuit (or channel unit) converts this to the idle code and sends it over the loop to the SCDSU as a pattern of seven bits with intentional bipolar violations. Likewise, the SCDSU sends the idle code over the loop to the line circuit. The line circuit converts the idle code to frames of seven bits at a logic-1 plus the eighth bit (least significant bit, LSB) set to logic-0; the latter indicates an idle control bit to the network.

• *SCDSU call origination*—Upon call origination, the SCDSU sends data mode, which is logic-1 in all seven bit positions, over the loop to the line circuit to indicate an off-hook. The line circuit then sends a logic-1 in bit-8 position (along with a logic-1 in the other seven bit positions) toward the network to indicate a service request. When the network is ready to receive address digits, it responds with a wink on the A-bit signaling channel. The wink is a momentary off-hook of normal duration, 140 to 290 ms. To indicate the wink, the network sends frames of eight bits at logic-1, although the eighth bit is a logic-1 only during the sixth frame. The network wink is translated by the line circuit to an all-ones wink toward the SCDSU. When the wink is completed, the line circuit returns to sending the idle code to the SCDSU.

• *Dialing*—The SCDSU dials by alternating between the data mode and idle using normal dial-pulse timing protocol. In response, the line circuit alternates between off-hook and on-hook toward the network in cadence with the dial pulses (eighth bit, or control bit, alternating between logic-1 and logic-0, all seven other bits set to logic-1). After dialing, the SCDSU returns to the data mode and sends a logic-1 in all seven bit positions to the line circuit. The line circuit, in turn, sends a logic-1 in all eight bit positions to the network. Until the called station answers, the originating line receives the idle code (seven bits of logic-1 plus bit-8 at logic-0) from the network and any call progress signaling toward the SCDSU is blocked.

• *Answer*—When the called station answers, the network sets the A-bit signaling channel to logic-1, which indicates a transition from idle to the data mode. The originating line circuit sends this over the loop as seven bit frames of all-ones. The SCDSU then sends data. Line circuits and channel units do not recognize analog call progress tones, such as ringback tone or busy tone. Unanswered calls cause the SCDSU to time-out after about 20 seconds.

TABLE VIII Interface Electrical Characteristics

Parameter	Type I	Type II	Type III
Line rate	56 kb/s	144 kb/s	160 kb/s
Interface	4-wire	2-wire	2-wire
Transmission method	Continuous	TCM	TCM
Line code	AMI (modified)	AMI	AMI
Transmitter signal level	1.69 v–p	2.73 v–p	2.55 v–p

B. SCDSU Call Termination

• *Idle*— The idle conditions described above also apply to this sequence.

• *Alerting*—The network alerts the terminating line circuit of a terminating call by going off-hook; that is, it sends data mode to the line circuit or channel unit. The terminating line circuit passes the data mode to the SCDSU as frames of all-ones.

• *Answer*—When the terminating DTE answers the call, the associated SCDSU sends all-ones to the terminating line circuit, which is a transition from idle to data mode. The line circuit passes this to the network as eight ones. The network then passes data between the originating and terminating SCDSUs.

VII. ISDN DIGITAL LINE

Developed in the 1980s, the ISDN digital line is the access, or subscriber loop, portion of the integrated services digital network (ISDN). Two interfaces are defined—144 kb/s for the basic rate interface (BRI) and 1.544 Mb/s (DS-1 rate) for the primary rate interface (PRI). From a physical layer

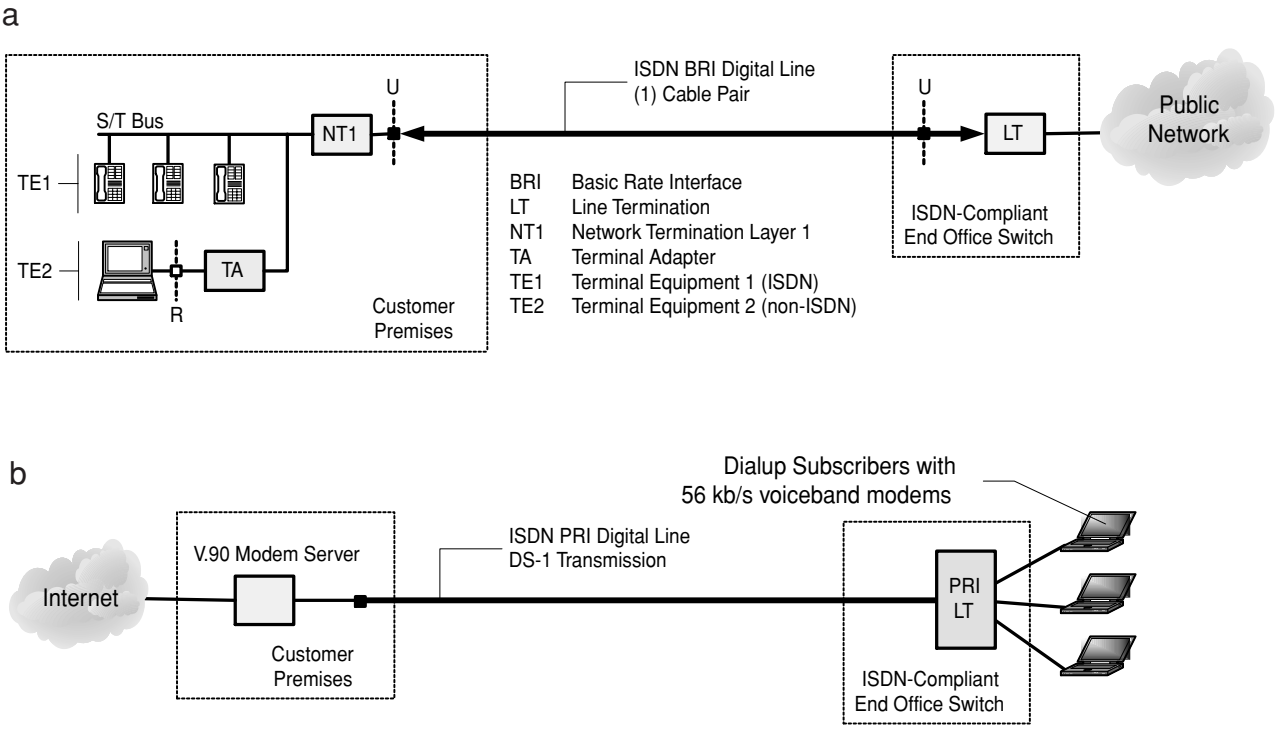


FIGURE 14 (a) The basic rate interface requires an ISDN-compliant, end-office switching system. The LT is an ISDN line card. The TA allows operation of analog and non-ISDN terminal equipment including PC serial ports and telephones. (b) The primary rate interface is used to connect Internet modem servers and ISDN-compliant PBXs to end-office switching systems.

perspective, the PRI is identical to the any other DS-1 rate interface. Typical interface applications are shown in Fig. 14.

Information is carried in the BRI using two full-duplex, circuit-switched B-channels and one full-duplex, packet-switched D-channel, all on a single cable pair. B-channels operate at 64 kb/s, and the D-channel in the BRI operates at 16 kb/s. B-channels may be combined to give the user higher composite rates for a particular application or call session. The BRI is somewhat unique when compared to earlier digital loop technologies, such as the subrate digital loops and repeatered T1-carrier in that the BRI was the first widely deployed digital loop to bidirectionally transmit data on only one cable pair.

Even though the B-channel is used mostly in the circuit switched mode (the B-channels are switched in an ISDN-compliant, end-office circuit switch), it also supports packet-switched data. The B-channel operates in a variety of modes including:

- Circuit switching for transparent 64 kb/s (clear channel) or nontransparent at lower rates.
- Packet switching, supporting packet mode terminals; in this case, the B-channel carries layers 2 and 3 of the OSI reference model according to, for example, International Telecommunications Union–Telecommunications Standardization Sector (ITU-T) recommendation X.25. The connection to the packet-switch network is first set up using a circuit switch.
- In both of the above modes, the connections through the circuit switch are considered semipermanent and exist only for the duration of the call session.

A B-channel is intended to carry a wide variety of user information. A distinguishing characteristic from conventional switched loops is that the B-channel does not carry signaling information for circuit switching in the ISDN. Signaling information used for circuit switching is carried by the D-channel. Typical information carried by the B-channel includes:

- Speech and analog voiceband data encoded at 64 kb/s using pulse code modulation (PCM)
- Wideband speech and music encoded using adaptive differential PCM (ADPCM) and a single 64-kb/s channel
- Speech encoded at rates lower than 64 kb/s alone, or combined with other digital information
- Digital data corresponding to circuit- or packet-switching user classes of service at bit rates equal to or lower than 64 kb/s

In the case of single information data streams at rates less than 64 kb/s, the data rate is adapted to be carried on the B-channel at 64 kb/s. Multiple streams may be multiplexed together in the same B-channel, but for circuit switching the entire B-channel is switched to a single user-network interface.

To support the various services, a number of B-channel connection types have been defined. Most simple ISDN terminal adapters (Fig. 14a) support three—data, speech, and 3.1-kHz audio. The data connection type is unrestricted in that the data bits are passed transparently through the network without any alteration. With 56 kb/s interworking, the data may have to be in a specific form, but this is a function of the terminal equipment. The speech connection type allows the greatest network flexibility and the least user flexibility. It allows echo cancellation and intermediate analog transmission, which will alter the data bit values in transit. It is intended primarily for digitally encoded speech, as its name indicates. The 3.1 kHz audio connection type is more restrictive on the network as far as how the data are manipulated and is more suitable for analog voiceband data than the speech connection type.

The BRI channel configuration is called 2B + D, giving a user payload rate of 144 kb/s (=2 × 64 kb/s + 16 kb/s). The local BRI loop actually operates at 160 kb/s, which includes the 144-kb/s payload plus an additional 16 kb/s of overhead for synchronization and framing.

The physical connection of the BRI at the subscriber's premises is through a U-interface, which uses a four-level pulse amplitude modulated (PAM) line code called 2B1Q, which stands for 2 binary, 1 quaternary. As user data are presented to the network termination layer 1 (NT1) at the interface, it is scrambled and interleaved with scrambled maintenance and operation channel (overhead) bits. The scrambled bits are combined with synchronization bits. Each pair of bits is then coded to a quaternary (quat) value, which corresponds to pulses with the voltage levels shown in Table IX. The quat value shown in this table is a reference value (analogous to binary-one or binary-zero) that is mapped to a voltage value. An example bit sequence and the resulting output from the 2B1Q line coder are shown in Fig. 15.

The basic pulse shape is the same for all quat values, the only difference being the polarity and pulse height (voltage) specified in the line code. Since the 2B1Q line code uses two bits per signal element (pulse), the line bit rate of 160 kb/s corresponds to a symbol rate of 80 kbaud. Each 2B1Q pulse has a width of 12.5 μs.

TABLE IX 2B1Q Line Code

Binary pair	Quat value	Pulse voltage (v–p)
10	+3	+2.5 (+15/6)
11	+1	+0.833 (+5/6)
01	−1	−0.833 (−5/6)
00	−3	−2.5 (−15/6)

The BRI originally was designed to be deployed without extensive engineering or qualification on any nonloaded loop meeting the design requirements for analog loops, generally a nonloaded 24-ga. pair less than 18,000 ft long (15,000 ft for 26-ga. cable). The BRI can share the cable with most analog loops and all baseband digital loops without cross-talk interference.

When the BRI is deployed, the existing twisted pair exchange cables will be used for most situations. A U-repeater, also called basic rate interface transmission extension (BRITE), may be used to extend the BRI to approximately twice its normal distance, or other technologies can be used to extend the loop to any distance within the limits of the transmission system.

Three fundamental technical advancements have allowed the BRI (and other current digital loop technologies) to be viable:

- Very large scale integration (VLSI)
- Decision feedback equalization (DFE) techniques
- Echo canceling

VLSI provides an economical packaging of the very complex electronic circuits, and DFE provides a means of reducing intersymbol interference caused by the imperfect transmission characteristics of twisted cable pairs. Echo cancellation allows full-duplex digital transmission on a single twisted pair. Economical DFE and echo canceling would not be possible without VLSI.

An echo canceller is required at each end of the 2-wire loop as shown in Fig. 16. The hybrid coupler performs the 2-wire to 4-wire conversion, but it does not keep all the transmitted signal out of the receiver (near echo) nor does it block the echoes that return from the hybrid couplers located at the far end of the circuit (far echo).

Adaptive transversal filters are used to generate replicas of the near- and far-end echoes, which are subtracted from the actual echoes. The filter adapts to the channel using a probing signal during the training sequence when the interfaces are connected to the loop and prior to actual user data transmission. The bulk delay preceding the far-end echo canceller is determined from a sounding of the channel or some equivalent adaptation procedure.

A variant of the BRI, ISDN digital subscriber line (ISDL), is used in applications such as dedicated Internet access. In this application, the BRI is not connected to an ISDN-compliant, end-office switching system but directly to a DSL access multiplexer (DSLAM). The two B-channels provide up to 128 kb/s of bandwidth; the

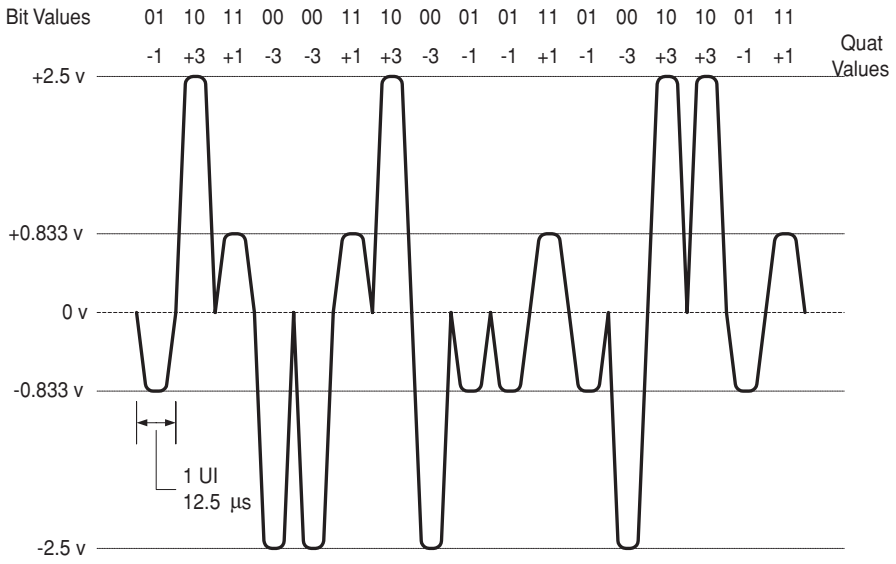


FIGURE 15 Example bit stream using the 2B1Q line code.

D-channel is not used for call setup (as in the BRI application), as no circuit-switching is involved.

While the BRI uses an interface designed specifically for 160-kb/s line speeds, the PRI uses existing interfaces operating at the DS-1 rate (1.544 Mb/s)—for example, repeatered T1-carrier, high-bit-rate digital subscriber line, or optical fiber with embedded DS-1 channels. The PRI generally is channelized as 23B + D, where B = 64 kb/s and D = 64 kb/s. The single D-channel handles signaling (for example, call setup, takedown, and management) associated with all 23 B-channels. As with all standard DS-1 rate interfaces, the user payload is 1.536 Mb/s. There is an additional 8 kb/s of overhead for framing and other network purposes, bringing the total rate to 1.544 Mb/s.

It is possible, depending on equipment features, to use a second PRI with 24B. In this case, the second PRI would use the D-channel of the first PRI for signaling purposes. This concept may be extended to additional PRI. The PRI is the preferred method to connect dialup subscribers to modem pools or servers located at Internet service providers (Fig. 14b). The PRI allows 64-kb/s clear channel transmission between the switch and the modem server, allowing analog voiceband 56-kb/s (V.90) modems to operate at faster speeds. (Without PRI, a DS-1 circuit setup for robbed bit signaling would be required and the individual channels in the DS-1 would be limited to 56 kb/s, thus limiting modem speed to something less than would be achievable on clear channels.)

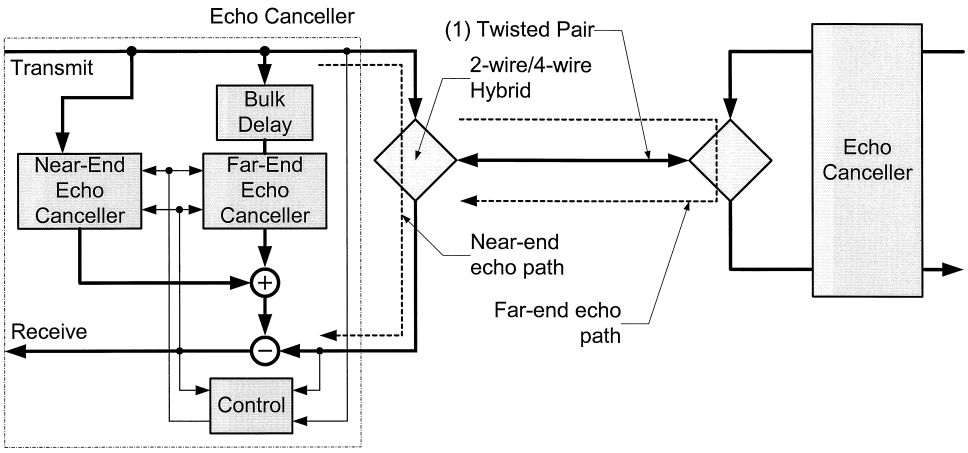


FIGURE 16 Echo canceller structure in an ISDN basic rate interface termination. The echo canceller subtracts a replica of the echo using two principal components—control mechanism and adaptive filters—to handle far-end and near-end echo.

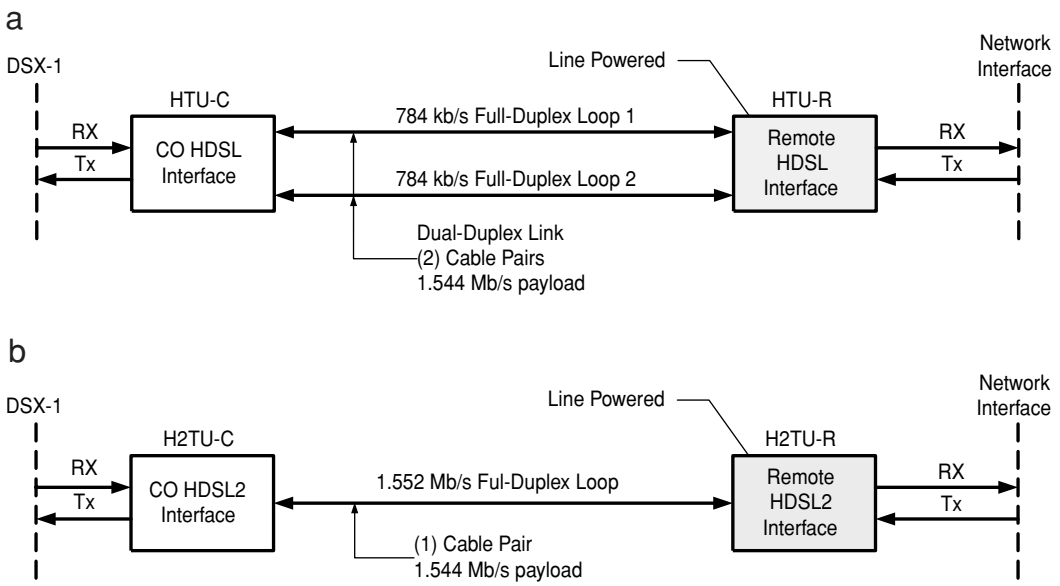


FIGURE 17 Both HDSL and HDSL2 transport DS-1 payload between two points over twisted pair cable. (a) HDSL; (b) HDSL2.

VIII. HIGH-BIT-RATE DIGITAL SUBSCRIBER LINE

The original high-bit-rate digital subscriber line (HDSL) was deployed in the public network in the early 1990s. It transports a user's DS-1 rate signal over two bidirectional twisted-pair loops, each operating at 784 kb/s. Using a technique called *dual-duplex transmission*, each of the two loops carries one-half of the total payload at 768 kb/s plus 8 kb/s DS-1 framing and 8 kb/s of additional HDSL overhead. The HDSL was conceived to allow quick deployment of DS-1 rate services without the high loop engineering and installation costs associated with repeatered T1-carrier. Unlike T1-carrier, HDSL can be used on loops with bridged taps, which are unterminated cable pairs connected to the active cable pairs used for HDSL service. The development of a standard for an improved version, called HDSL2, was completed in the late 1990s. The HDSL2, or second-generation HDSL, requires only one cable pair but delivers the same DS-1 rate interface as HDSL. Both HDSL and HDSL2 are used in point-to-point applications only.

The HDSL originally was designed as a repeaterless technology but manufacturers developed line repeaters to double and triple the loop length from the original maximum of 12,000 ft to as much as 24,000 and 36,000 ft. The HDSL2 was designed from the outset to use one or two line repeaters to extend the range up to 36,000 ft. If any 26-ga. cable is present on HDSL loops, the range is reduced to as low as 9,000 ft per repeatered segment.

Figure 17 shows how the HDSL terminal units are implemented. Figure 18 shows a typical application of the HDSL. HDSL2 may be used in the same applications as HDSL but the former requires only one-half the outside plant to accomplish the same job. The HDSL remote terminal (HTU-R) normally is line powered by the

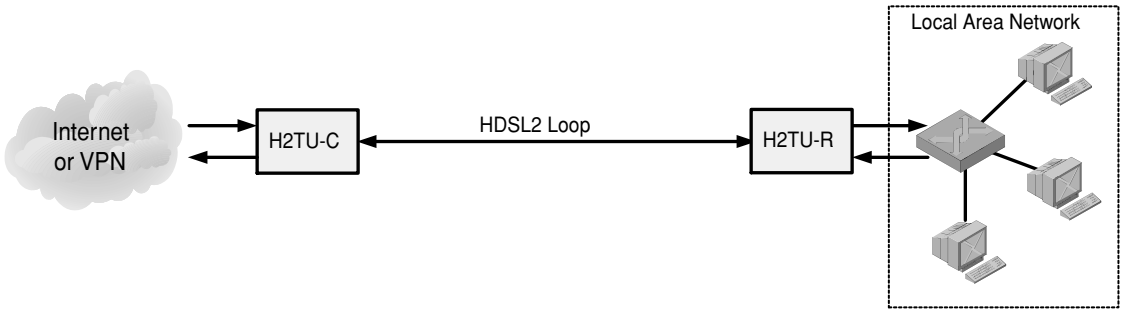


FIGURE 18 A typical HDSL application provides access to the Internet or a virtual private network (VPN).

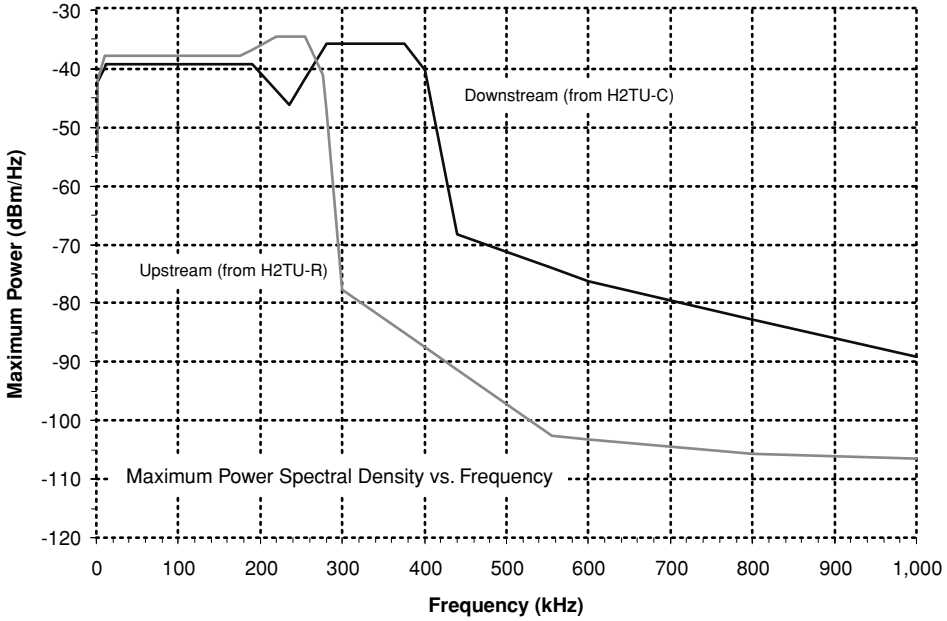


FIGURE 19 HDSL2 power spectral density masks. The PSD control technique used by HDSL2 and SHDSL generally is referred to as overlapped PAM transmission with interlocking spectra, or OPTIS.

HDSL central office terminal (HTU-C). On the network side, the HTU-C provides a standard signal compatible with existing digital signal cross-connect DSX-1 cross-connect equipment in central offices and is thus transparent to the network. The HTU-R can be programmed to provide a DSX-1 compatible signal or a signal that complies with industry standards for user's digital signal DS-1 rate network interface. The latter uses a slightly different pulse template that recognizes the different types of cables found on user premises. In the HDSL2, the two terminals are designated H2TU-C and H2TU-R.

The HDSL loop interface uses the 2B1Q line code, which is identical to that used on the ISDN digital line already described. As user data are presented to the HTU-C or HTU-R, they are scrambled and interleaved with maintenance and operation channel (overhead) bits and then combined with synchronization bits as with the ISDN digital line.

The shape and amplitude of the pulses at the loop interface are identical to the ISDN BRI digital line, but the pulse width is different to accommodate the higher signaling rate. In the case of the HDSL, the pulse width is specified in a unit interval of 2.55 μs. The basic pulse shape is the same for all quat values, the only difference being the polarity and pulse voltage. The HDSL line rate of 784 kb/s on each of the two loops corresponds to a symbol rate of 392 kbaud. Since each of the two HDSL loops is bidirectional, 2-wire to 4-wire hybrid couplers and echo cancellers are required to keep the transmitted signal from interfering with the received signal. The techniques are identical to those described for the ISDN BRI digital line.

The HDSL2 digital loop also uses echo cancellers and hybrid couplers; however, it is unique in its application of a one-dimensional, trellis-coded, 16-level pulse amplitude modulated (TC-PAM) line code. HDSL2 pulse characteristics are controlled by controlling the power spectral density (PSD) of the transmitted signals using programmable filters. Even though the HDSL2 provides symmetrical transmission, the upstream and downstream PSD are different but do overlap (Fig. 19). Controlled PSD is necessary to control cross-talk interference between digital loops sharing the same twisted-pair cable. Spectrum compatibility and cross-talk control are fundamentally important to the compatibility of new and old loop technologies. The basic TC-PAM techniques also have been adopted for the SHDSL, which is a higher speed, rate-adaptable version of HDSL2, but the PSD masks are slightly different. The modulation techniques used by HDSL2 result in much more complex interface equipment (integrated chipset) than the original HDSL.

Trellis encoding is a form of forward error correction (FEC) that convolutionally encodes the transmitted signal points (the term "trellis" refers to the appearance of the path diagram that depicts coder state changes). A trellis coding algorithm adds redundancy by accepting m data bits as input and generating $m+1$ bits as output. It is useful in bandwidth-limited channels because it provides

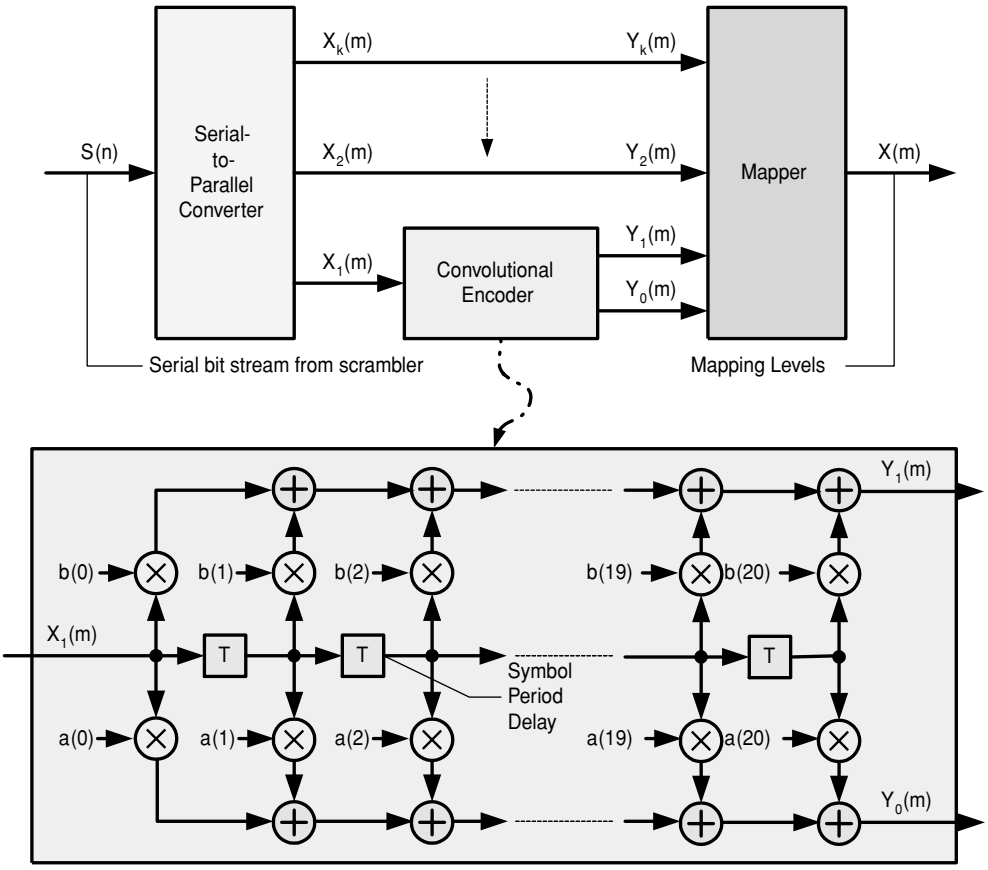


FIGURE 20 Trellis encoder and mapper for HDSL2. The binary coefficients *a* and *b* in the programmable convolutional encoder are determined during the activation phase when the H2TU-C and H2TU-R are first connected to the loop. They determine which taps in the encoder shift register are used to form the output and thus determine the number of states and specific trellis code.

FEC without increasing the bandwidth requirements. In the case of HDSL2, three data bits are input to the trellis encoder and four bits are output (Fig. 20).

Trellis encoding combines modulation/FEC encoding and demodulation/FEC decoding in an integral operation that effectively increases the minimum distance between the signals that are most likely to be confused in the decoder at the other end of the loop. Sequences of trellis-coded symbols are restricted to certain valid patterns. These patterns are chosen to have a large minimum distance between them so that corrupted sequences can be corrected, thus reducing the error probability of the decoded signals.

The four bits from the trellis encoder are mapped to one of 16 specific levels by the PAM line coder as shown in Table X. Each mapping level corresponds to one of 16 PSD masks (numbered 0 to 15) separated by 1 dB. The total power in mask 0 is approximately +16.5 dBm, corresponding to a voltage level of about 2.5 volts,

TABLE X HDSL2 Line Code Mapping

Binary bit groups	Mapping level $x(m)$
0000	−15/16
0001	−13/16
0010	−11/16
0011	−9/16
0100	−7/16
0101	−5/16
0110	−3/16
0111	−1/16
1100	+1/16
1101	+3/16
1110	+5/16
1111	+7/16
1000	+9/16
1001	+11/16
1010	+13/16
1011	+15/16

and in mask 15 it is +1.5 dBm, corresponding to a voltage level of about 437 mv. The HDSL2 line rate is 1.552 Mb/s and the symbol rate is $1552 \div 3 =$ 517.33 kb/s.

A high internal digital signal processing load, as with HDSL2 and SHDSL, increases the latency of loop transmission. Latency, as it applies to digital loops, is the delay incurred between the time a user bit enters the DSL interface on one end and when it exits from the associated DSL interface at the other end. All digital network nodes add to latency as user data are transported on an end-to-end basis. Real-time applications, such as speech and live video, are very sensitive to delay. The HDSL2 specifies no more than 500-μs latency, and the SHDSL specifies no more than 1.2 μs; the latter recognizes the additional signal processing of SHDSL. While these values appear small, they add to the latency of many other nodes in any end-to-end digital connection. Speech is particularly sensitive to delay if there is echo present. Delayed echo makes conversation difficult and becomes annoying if the roundtrip delay exceeds approximately 8 ms.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER NETWORKS • DIGITAL CONTROLLERS • DIGITAL ELECTRONIC CIRCUITS • DIGITAL FILTERS • NETWORKS FOR DATA COMMUNICATION • WIRELESS COMMUNICATIONS

BIBLIOGRAPHY

Bingham, J. A. C. (2000). "ADSL, VDSL, and Multicarrier Modulation," John Wiley & Sons, New York.

Chen, W. Y. (1998). "DSL Simulation Techniques and Standards Development for Digital Subscriber Lines," Macmillan, New York.

Reeve, W. D. (1995). "Subscriber Loop Signaling and Transmission Handbook—Digital," IEEE Press, New York.

Reeve, W. D. (1998). "Subscriber Loop Signaling." *In* "The Froehlich/Kent Encyclopedia of Telecommunications," Vol. 16, Marcel Dekker, New York.

Rauschmayer, D. J. (1999) "ADSL/VDSL Principles," Macmillan, New York.

Sklar, B. (2000). "Digital Communications—Fundamentals and Applications," Prentice Hall, New York.

Starr, T., Cioffi, J., and Silverman, P. (1999). "Understanding Digital Subscriber Line Technology," Prentice Hall, New York.



Communication Systems, Civilian

Simon Haykin
McMaster University

I. Model of a Communication System
II. Analog Modulation
III. Pulse and Digital Modulation
IV. Data Transmission
V. Spread-Spectrum Modulation
VI. Multiplexing

GLOSSARY

Amplitude modulation (AM) Process in which the amplitude of a carrier is varied about a mean value, linearly with a baseband signal.

Average probability of symbol error Probability that the symbol recreated at the receiver output of a digital communication system differs from that transmitted, on the average.

Baseband Bank of frequencies representing an original signal as delivered by a source of information.

Bit Acronym for binary digit.

Digital subscriber line (DSL) A system that connects a user terminal (e.g., computer) to a telephone company's central office.

Discrete multitone (DMT) A form of digital modulation using the discrete Fourier transform.

Double-sideband suppressed-carrier (DSBSC) modulation Modulated wave resulting from multiplying a baseband signal by carrier wave.

Frequency modulation (FM) Form of angle modulation in which the instantaneous frequency is varied linearly with the baseband signal.

Frequency-shift keying (FSK) Digital modulation in which the frequency of a carrier takes on a discrete value chosen from a pre-selected set in accordance with the input data.

Orthogonal frequency-division multiplexing (OFDM) Another form of multichannel modulation that differs from DMT in areas of application and some aspects of design.

Phase modulation (PM) Form of angle modulation in which the angle of a carrier is varied linearly with the baseband signal.

Phase-shift keying (PSK) Digital modulation in which the phase of a carrier takes on a discrete value chosen from a preselected set in accordance with the input data.

Prediction Filtering operation used to make an inference about the future value of a baseband signal, given knowledge of past behavior of the signal up to a certain point in time.

Pulse-analog modulation Modulation process in which some characteristic feature (e.g., amplitude, duration, or position) of each pulse in a periodic pulse train (used as carrier) is varied in accordance with the baseband signal.

Pulse-code modulation (PCM) Process that involves the uniform sampling of a baseband signal, quantization of the amplitude of each sample to the nearest one of a finite set of allowable values, and representation of each quantized sample by a code.

Signal-to-noise ratio Ratio of the average power of a message signal to the average power of additive noise, both being measured at a point of interest in the system (e.g., receiver output).

Single-sideband (SSB) modulation Form of amplitude modulation in which only the upper sideband or lower sideband is transmitted.

Vestigial sideband (VSB) modulation Form of amplitude modulation in which one sideband is passed almost completely, whereas just a vestige, or trace, of the other sideband is retained.

I. MODEL OF A COMMUNICATION SYSTEM

Figure 1 shows the block diagram of a communication system. The system consists of three major parts: (1) transmitter, (2) communication channel, and (3) receiver. The main purpose of the transmitter is to change the message signal into a form suitable for transmission over the channel. This modification is achieved by means of a process known as modulation.

The communication channel may be a transmission line (as in telephony and telegraphy), an optical fiber (as in optical communications), or merely free space in which the signal is radiated as an electromagnetic wave (as in wireless communications and radio and television broadcasting). In propagating through the channel, the transmitted signal is distorted because of nonlinearities or imperfections in the frequency response of the channel. Other sources of degradation are noise and interference picked up by the signal during the course of transmission through the channel. Noise and distortion constitute two basic problems in the design of communication systems. Usually, the transmitter and receiver are carefully designed so as to minimize the effects of noise and distortion on the quality of reception.

The main purpose of the receiver is to recreate the original message signal from the degraded version of the transmitted signal after propagation through the channel. This recreation is accomplished by a process known as demodulation, which is the reverse of the modulation process in the transmitter. However, owing to the unavoidable presence of noise and distortion in the received signal, the receiver cannot recreate the original message signal exactly. The resulting degradation in overall system performance is influenced by channel characteristics and the type of modulation scheme used. Some modulation schemes are less sensitive to the effects of noise and distortion than others.

In any communication system, there are two primary communication resources, namely, transmitted power and channel bandwidth. A general system design objective is to use these two resources as efficiently as possible. In most communication channels, one resource may be considered more important than the other. We therefore classify communication channels as power-limited or band-limited. For example, a telephone circuit is a typical band-limited channel, whereas a space communication link or a satellite channel is typically power-limited.

When the spectrum of a message signal extends down to zero or low frequencies, we define the bandwidth of the signal as that upper frequency above which the spectral content of the signal is negligible and, therefore, unnecessary for transmitting the pertinent information. For example, the average voice spectrum extends well beyond 10 kHz, though most of the energy is concentrated in the range of 100 to 600 Hz, and a band from 300 to 3400 Hz gives good articulation. Accordingly, telephone circuits that respond well to the latter range of frequencies give quite satisfactory commercial telephone service.

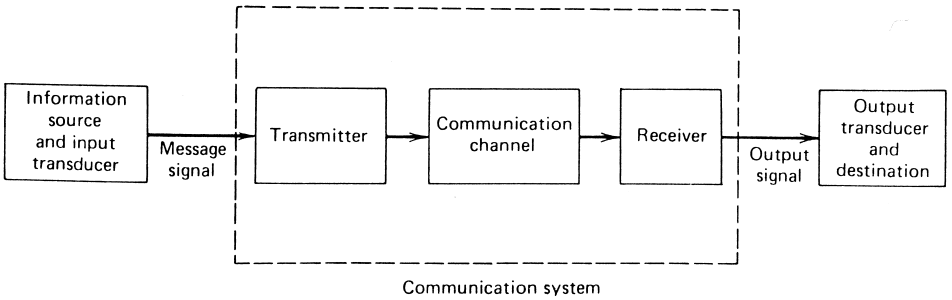


FIGURE 1 Model of an electrical communication system. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

II. ANALOG MODULATION

Modulation is formally defined as the process by which some characteristic of a carrier is varied in accordance with a modulating wave. The baseband signal is referred to as the modulating wave, and the result of the modulation process is referred to as the modulated wave.

In analog or continuous-wave (cw) modulation, the modulating wave consists of an analog signal (e.g., voice signal, video signal), and the carrier consists of a sine wave. Basically, there are two types of analog modulation: amplitude modulation and angle modulation. In amplitude modulation the amplitude of the sinusoidal carrier wave is varied in accordance with the baseband signal. In angle modulation, on the other hand, the angle of the sinusoidal carrier wave is varied in accordance with the baseband signal.

A. Amplitude Modulation

Consider a sinusoidal carrier wave $c(t)$ defined by

$$c(t) = A_c \cos(2\pi f_c t) \tag{1}$$

where A_c is the carrier amplitude and f_c the carrier frequency. For convenience, we have assumed that the phase of the carrier wave is zero in Eq. (1). Let $m(t)$ denote the baseband signal that carries the specification of the message. The carrier wave $c(t)$ is independent of $m(t)$. An amplitude-modulated (AM) wave is described by

$$s(t) = A_c[1 + k_a m(t)] \cos(2\pi f_c t) \tag{2}$$

where k_a is a constant called the amplitude sensitivity of the modulator.

The envelope of the AM wave $s(t)$ equals

$$a(t) = A_c|1 + k_a m(t)| \tag{3}$$

where $|\cdot|$ denotes the absolute value of the enclosed quantity. The envelope of $s(t)$ has the same shape as the baseband signal $M(t)$, provided that two requirements are satisfied:

1. The amplitude of $k_a m(t)$ is always less than unity, that is,

$$|k_a m(t)| < 1 \quad \text{for all } t$$

The absolute maximum value of $k_a m(t)$ multiplied by 100 is referred to as the percentage modulation.

2. The carrier frequency f_c is much greater than the highest frequency component of the baseband signal $m(t)$.

The spectrum of the AM wave $s(t)$ consists of a carrier, upper sideband, and lower sideband. Let W denote the highest frequency component of the baseband signal $m(t)$. For positive frequencies, the carrier is located at f_c, the upper sideband extends from f_c to $f_c + W$, and the lower sideband extends from $f_c - W$ to f_c. Hence, the transmission bandwidth of the AM wave equals $2W$, that is, exactly twice the message bandwidth. For negative frequencies, the spectrum of the AM wave is the mirror image of that for positive frequencies.

Figure 2 illustrates the amplitude modulation process. Figure 2a shows a modulating wave $m(t)$ that consists of a single tone or frequency component, Fig. 2b the sinusoidal carrier wave, and Fig. 2c the corresponding AM wave. The amplitude spectra of the modulating wave, the carrier, and the AM wave are shown on the right-hand side of the figure.

1. Double-Sideband Suppressed-Carrier Modulation

The carrier wave $c(t)$ is completely independent of the information-carrying signal or baseband signal $m(t)$, which means that the transmission of the carrier wave represents a waste of power. This points to a shortcoming of amplitude modulation, namely, that only a fraction of the total transmitted power is affected by $m(t)$. To overcome this shortcoming, we may suppress the carrier component from the modulated wave, resulting in double-sideband suppressed-carrier (DSBSC) modulation. Thus, by suppressing the carrier, we obtain a modulated wave that is proportional to the product of the carrier wave and the baseband signal.

To describe a DSBSC-modulated wave as a function of time, we write

$$s(t) = c(t)m(t) = A_c \cos(2\pi f_c t)m(t) \tag{4}$$

This modulated wave undergoes a phase reversal whenever the baseband signal $m(t)$ crosses zero. Accordingly, unlike amplitude modulation, the envelope of a DSBSC-modulated wave is different from the baseband signal. For obvious reasons, a device that performs the operation described in Eq. (4) is called a product modulator.

The transmission bandwidth required by DSBSC modulation is the same as that for amplitude modulation, namely, $2W$. Figure 2d illustrates the DSBSC-modulated wave for single-tone modulation and the corresponding amplitude spectrum.

2. Quadrature-Carrier Multiplexing

The quadrature-carrier multiplexing or quadrature-amplitude modulation (QAM) scheme enables two DSBSC-modulated waves (resulting from the application of two independent message signals) to occupy the same

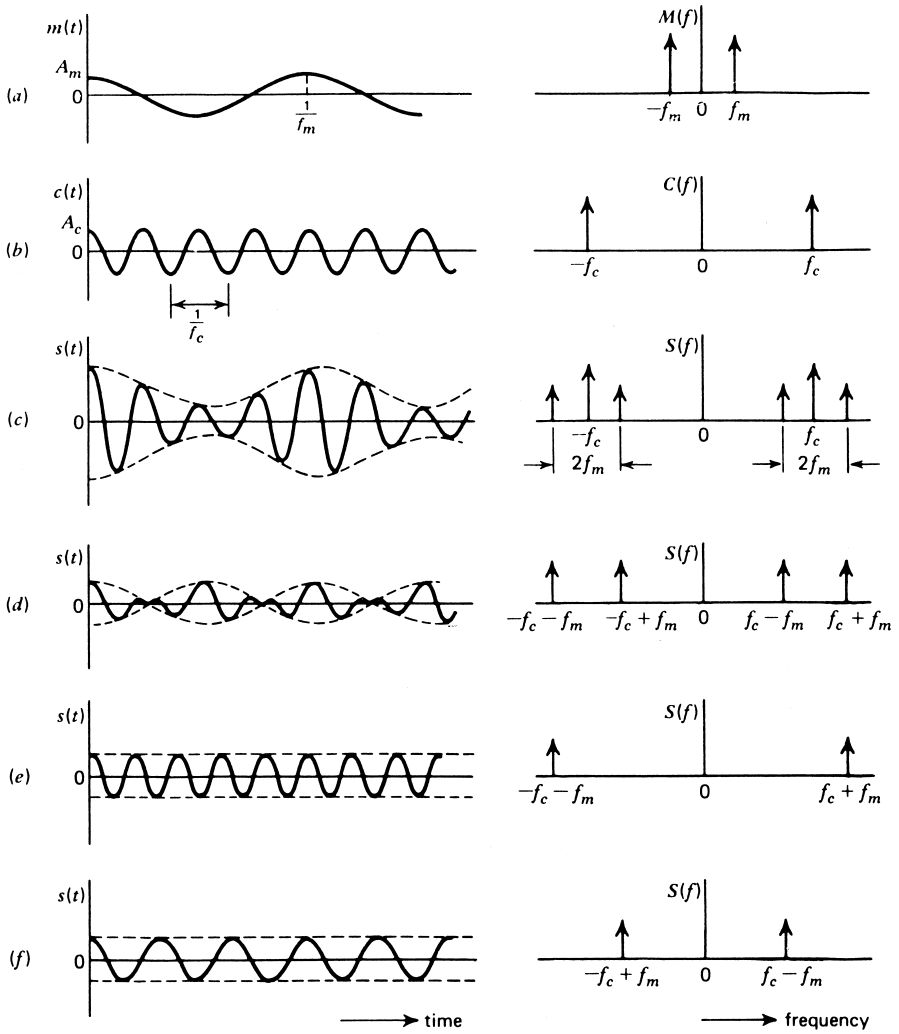


FIGURE 2 Time-domain (left) and frequency-domain (right) characteristics of different modulated waves produced by a single tone. (a) Modulating wave. (b) Carrier wave. (c) AM wave. (d) DSBSC wave. (e) SSB wave with the upper-side frequency transmitted. (f) SSB wave with the lower-side frequency transmitted. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

transmission bandwidth, and yet it allows for the separation of the two message signals at the receiver output. It is therefore a bandwidth-conservation scheme. The transmitter part of the system involves the use of two separate product modulators that are supplied with two carrier waves of the same frequency but differing in phase by 90°. The transmitted signal $s(t)$ consists of the sum of these two product modulator outputs, as shown by

$$s(t) = A_c m_1(t)\cos(2\pi f_c t) + A_c m_2(t)\sin(2\pi f_c t) \quad (5)$$

where $m_1(t)$ and $m_2(t)$ denote the two different message signals applied to the product modulators. Thus, $s(t)$ occupies a transmission bandwidth of $2W$, centered at the carrier frequency f_c, where W is the message bandwidth of $m(t)$ or $m_2(t)$.

Quadrature-carrier multiplexing is used in color television. In this system, synchronizing pulses are transmitted in order to maintain the local oscillator in the receiver at the correct frequency and phase with respect to the carrier used in the transmitter.

3. Single-Sideband Modulation

Amplitude modulation and DSBSC modulation are wasteful of bandwidth because they both require a transmission bandwidth equal to twice the message bandwidth. In either case, one-half of the transmission bandwidth is occupied by the upper sideband of the modulated wave, whereas the other half is occupied by the lower sideband. However, the upper and lower sidebands are uniquely related

to each other by virtue of their symmetry about the carrier frequency; that is, given the amplitude and phase spectra of either sideband, we can uniquely determine the other. This means that insofar as the transmission of information is concerned, only one sideband is necessary, and if the carrier and the other sidebands are suppressed at the transmitter, no information is lost. Thus, the communication channel has to provide only the same bandwidth as the baseband signal, a conclusion that is intuitively satisfying. When only one sideband is transmitted, the modulation system is referred to as a single-sideband (SSB) system.

The precise frequency–domain description of an SSB wave depends on which sideband is transmitted. In any event, the essential function of an SSB modulation system is to translate the spectrum of the modulating wave, either with or without inversion, to a new location in the frequency domain; the transmission bandwidth requirement of the system is one-half that of an amplitude or DSBSC modulation system. The benefit of using SSB modulation is therefore derived principally from the reduced bandwidth requirement and the elimination of the high-power carrier wave. The principal disadvantage of the SSB modulation system is its cost and complexity.

The time-domain description of the SSB wave is much more complicated than that of the DSBSC wave. Specifically, we write

$$s(t) = \tfrac{1}{2}A_c m(t)\cos(2\pi f_c t) \pm \tfrac{1}{2}A_c \hat{m}(t)\sin(2\pi f_c t) \quad (6)$$

where $\hat{m}(t)$ is the Hilbert transform of the baseband signal $m(t)$. A Hilbert transformer consists of a two-port device that leaves the amplitudes of all frequency components of the input signal unchanged, but it produces a phase shift of -90° for all positive frequency components of the input signal and a phase shift of $+90^\circ$ for all negative frequency components. The minus sign on the right-hand side of Eq. (6) refers to an SSB-modulated wave that contains only the upper sideband, whereas the plus sign refers to an SSB-modulated wave that contains only the lower sideband.

Figure 2e shows the SSB wave (with upper-side frequency) resulting from the use of single-tone modulation. Figure 2f shows the corresponding SSB wave with lower-side frequency.

4. Vestigial Sideband Modulation

Single-sideband modulation is rather well suited for the transmission of voice because of the energy gap in the spectrum of voice signals between zero and a few hundred hertz. When the baseband signal contains significant components at extremely low frequencies (as in the case of television and telegraph signals and computer data), however, the upper and lower side-bands meet at the carrier frequency. This means that the use of SSB modulation is inappropriate for the transmission of such baseband signals because of the difficulty of isolating one sideband. This difficulty suggests another scheme known as vestigial sideband (VSB) modulation, which is a compromise between SSB and DSBSC modulation. In this modulation scheme, one side-band is passed almost completely, whereas just a trace, or vestige, of the other sideband is retained. The transmission bandwidth required by a VSB system is therefore

$$B_T = W + f_v \quad (7)$$

where W is the message bandwidth and f_v the width of the vestigial sideband.

Vestigial sideband modulation has the virtue of conserving bandwidth almost as efficiently as SSB modulation, while retaining the excellent low-frequency baseband characteristics of double-sideband modulation. Thus, VSB modulation has become standard for the transmission of television and similar signals where good phase characteristics and transmission of low-frequency components are important but the bandwidth required for double-sideband transmission is unavailable or uneconomical.

It is of interest, however, that in commercial television broadcasting the transmitted signal is not quite VSB-modulated, because the shape of the transition region is not rigidly controlled at the transmitter. Instead, a VSB filter is inserted in each receiver. The overall performance is the same as VSB modulation except for some wasted power and bandwidth. Figure 3 shows the idealized frequency response for a VSB filter designed for television receivers.

B. Angle Modulation

In angle modulation the angle of the carrier wave is varied according to the baseband signal. In this method of modulation the amplitude of the carrier wave is maintained constant. An important feature of angle modulation is that it can provide better discrimination against noise and interference than amplitude modulation. However, this improvement in performance is achieved at the expense of increased transmission bandwidth; that is, angle modulation provides a practical means of exchanging transmission bandwidth for improved noise performance. Such a trade-off is not possible with amplitude modulation.

Two forms of angle modulation are distinguished: phase modulation and frequency modulation. These two methods of modulation are closely related in that the properties of one can be derived from those of the other. It is customary to define these two forms of angle modulation as follows:

1. Phase modulation (PM) is that form of angle modulation in which the angle $\theta_i(t)$ is varied linearly with the baseband signal $m(t)$, as shown by

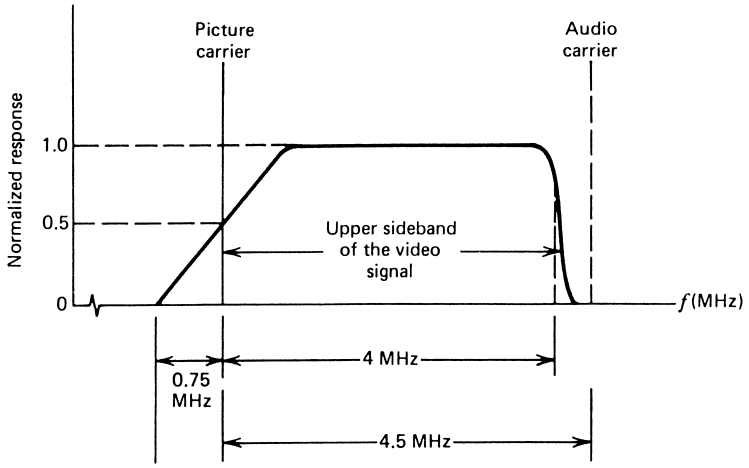


FIGURE 3 Frequency response of a VSB filter used in television receivers. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

$$\theta_i(t) = 2\pi f_c t + k_p m(t) \tag{8}$$

The term $2\pi f_c t$ represents the angle of the unmodulated carrier, and the constant k_p represents the phase sensitivity of the modulator. For convenience, we have assumed in Eq. (8) that the angle of the unmodulated carrier is zero at $t = 0$. The PM wave $s(t)$ is thus described in the time domain by

$$s(t) = A_c \cos[2\pi f_c t + k_p m(t)] \tag{9}$$

2. Frequency modulation (FM) is that form of angle modulation in which the instantaneous frequency $f_i(t)$ is varied linearly with the baseband signal $m(t)$, as shown by

$$f_i(t) = \frac{1}{2\pi}\frac{d_i(t)}{dt} = f_c + k_f m(t) \tag{10}$$

The term f_c represents the frequency of the unmodulated carrier, and the constant k_f represents the frequency sensitivity of the modulator. Integrating Eq. (10) with respect to time and multiplying the result by 2π, we get

$$\theta_i(t) = 2\pi f_c t + 2\pi k_f \int_0^t m(t)\,dt \tag{11}$$

where, for convenience, we have assumed that the angle of the unmodulated carrier wave is zero at $t = 0$. The FM is therefore described in the time domain by

$$s(t) = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)\,dt\right] \tag{12}$$

A consequence of allowing the angle $\theta_i(t)$ to become dependent on the message signal $m(t)$ as in Eq. (8) or on its integral as in Eq. (11) is that the zero crossings of PM wave or FM wave no longer have a perfect regularity in their spacing; zero crossings refer to the instants of time at which a waveform changes from negative to positive value or vice versa.

The differences between amplitude-modulated and angle-modulated waves are illustrated in Fig. 4 for the case of single-tone modulation. Figures 4a and b refer to the sinusoidal carrier and modulating waves, respectively; Figs. 4c, d, and e show the corresponding AM, PM, and FM waves, respectively.

Figure 5a shows that an FM wave can be generated by first integrating $m(t)$ and then using the result as the input to a phase modulator. Conversely, a PM wave can be generated by first differentiating $m(t)$ and then using the result as the input to a frequency modulator, as in Fig. 5b.

Consider an FM wave produced by using a single-tone modulating wave of frequency f_m. Let Δf denote the frequency deviation defined as the maximum departure of the instantaneous frequency of the FM wave from the carrier frequency f_c. The ratio $\Delta f/f_m$ is called the modulation index, commonly denoted by β. When β is less than 0.5, the modulated wave is referred to as narrow-band FM. When β is large, it is referred to as wide-band FM.

1. Transmission Bandwidth of FM Waves

In theory, an FM wave contains an infinite number of side frequencies so that the bandwidth required to transmit such a signal is similarly infinite in extent. In practice, however, the FM wave is effectively limited to a finite number of significant side frequencies compatible with a specified amount of distortion. We can therefore specify an effective bandwidth required for the transmission of an FM wave.

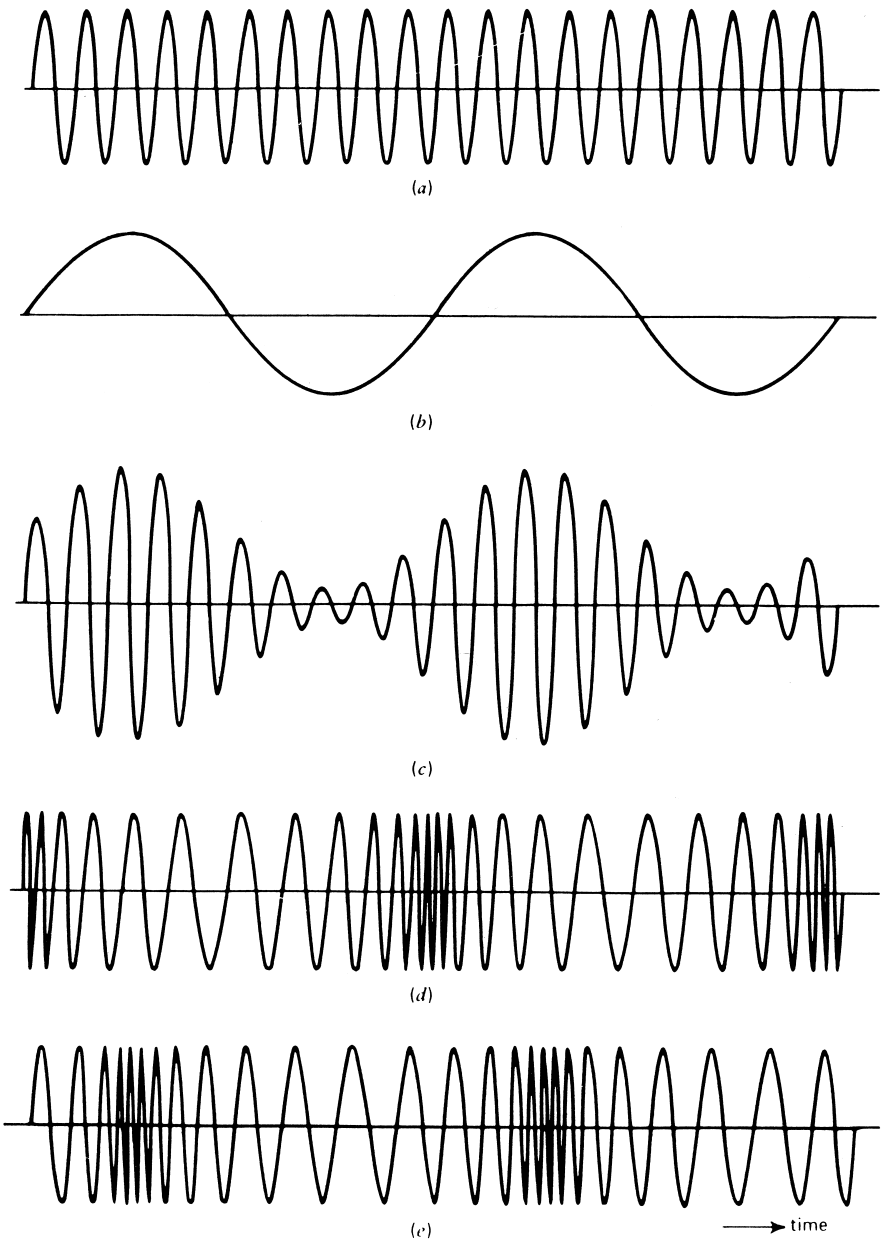


FIGURE 4 AM, PM, and FM waves produced by a single tone. (a) Carrier wave. (b) Sinusoidal modulating wave. (c) AM wave. (d) PM wave. (e) FM wave. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

For the case of single-tone modulation, we calculate the transmission bandwidth of the FM wave by using Carson's rule:

$$B_T \backsimeq 2\Delta f + 2f_m = 2\Delta f(1 + 1/\beta) \qquad (13)$$

When β is small compared with 1 rad, as in the case of narrow-band FM, the transmission bandwidth B_T is approximately $2f_m$. On the other hand, in the case of wide-band FM for which β is large compared with 1 rad, the transmission bandwidth B_T is approximately $2\Delta f$.

C. Effects of Noise

One of the basic issues in the study of modulation systems is the analysis of the effects of noise on the performance of the receiver and the use of the results of this analysis in system design. Another matter of interest is the comparison of the noise performances of different modulation–demodulation schemes. In order to carry out this analysis, we obviously need a criterion that describes in a meaningful manner the noise performance of the modulation

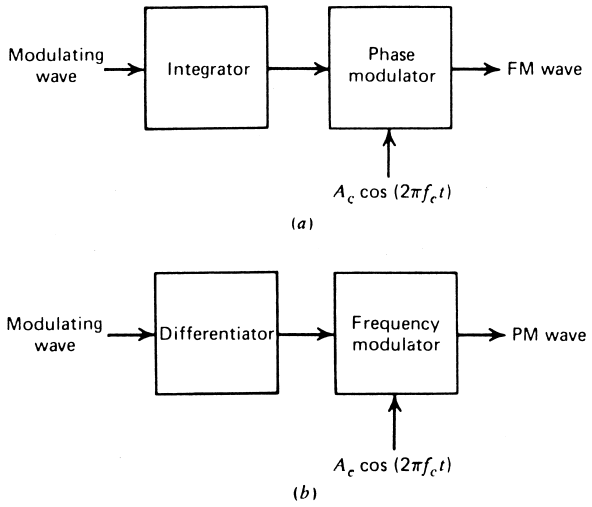


FIGURE 5 Relationship between frequency modulation and phase modulation. (a) Scheme for generating an FM wave by using a phase modulator. (b) Scheme for generating a PM wave by using a frequency modulator. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

system under study. In the case of cw modulation systems, the customary practice is to use the output signal-to-noise ratio as an intuitive measure for describing the fidelity with which the demodulation process in the receiver recovers the original message from the received modulated signal in the presence of noise. The output signal-to-noise ratio is defined as the ratio of the average power of the message signal to the average power of the noise, both measured at the receiver output. Such a criterion is perfectly well defined as long as the message signal and noise appear additively at the receiver output.

In performing noise analysis, it is customary to model the noise as white, Gaussian, and stationary. Under this condition, we find that for the same average transmitted (or modulated message) signal power and the same average noise power in the message bandwidth, an SSB receiver will have exactly the same output signal-to-noise ratio as a DSBSC receiver, when both receivers use coherent detection for the recovery of the message signal. Furthermore, in both cases, the noise performance of the receiver is the same as that obtained by simply transmitting the message signal itself in the presence of the same noise. The only effect of the modulation process is to translate the message signal to a different frequency band.

Coherent detection assumes the availability of a local carrier in the receiver with the same frequency and phase as that of the carrier in the transmitter. The noise performance of an AM receiver is always inferior to that of a DSBSC or SSB receiver. This is due to the wastage of transmitter power that results from transmitting the carrier as a component of the AM wave. Also, an AM receiver suffers from a threshold effect that arises when the carrier-to-noise ratio is small compared with unity. By threshold we mean a value of the carrier-to-noise ratio below which the noise performance of a detector deteriorates much more rapidly than proportionately to the carrier-to-noise ratio. A detailed analysis of the threshold effect in envelope detectors is complicated, however.

Since frequency modulation is a nonlinear modulation process, the noise analysis of an FM receiver is also quite complicated. Nevertheless, we can state that when the carrier-to-noise ratio is high, an increase in the transmission bandwidth B_T of the FM system provides a corresponding quadratic increase in the output signal-to-noise ratio of the system. However, this result is valid only if the carrier-to-noise ratio is high compared with unity. It is found experimentally that as the input noise is increased so that the carrier-to-noise ratio is decreased, the FM receiver breaks. At first, individual clicks are heard in the receiver output, and as the carrier-to-noise ratio decreases still further, the clicks rapidly merge into a crackling or sputtering sound. Under this condition, the FM receiver is said to suffer from the threshold effect.

1. Preemphasis and Deemphasis

The noise performance of an FM receiver can be improved significantly by the use of preemphasis in the transmitter and deemphasis in the receiver. With this method, we artificially emphasize the high-frequency components of the message signal before modulation in the transmitter and therefore before the noise is introduced in the receiver. In effect, the low-frequency and high-frequency portions of the power spectral density of the message are equalized in such a way that the message fully occupies the frequency band allotted to it. Then, at the discriminator output in the receiver, we perform the inverse operation by deemphasizing the high-frequency components, so as to restore the original signal-power distribution of the message. In such a process the high-frequency components of the noise at the discriminator output are also reduced, thereby effectively increasing the output signal-to-noise ratio of the system.

The use of the simple linear preemphasis and deemphasis filters is an example of how the performance of an FM system can be improved by utilizing the differences between characteristics of signals and noise in the system. These simple filters also find application in audio tape recording. In recent years nonlinear preemphasis and deemphasis techniques have been applied successfully to tape recording. These techniques (known as the Dolby-A, Dolby-B, and DBX systems) use a combination of filtering and dynamic range compression to reduce

the effects of noise, particularly when the signal level is low.

2. Noise Comparison of Analog Modulation Schemes

Comparing the noise performance of different CW modulation schemes, we may make the following observations:

1. Among the family of AM schemes, SSB modulation is optimum.
2. The use of FM improves noise performance, but at the expense of an increased transmission bandwidth, assuming that the FM system operates above threshold for the noise improvement to be realized. The exchange of increased transmission bandwidth for improved noise performance follows a square law, which is the best that can be done with cw modulation.

III. PULSE AND DIGITAL MODULATION

In cw modulation, some parameter of a sinusoidal carrier wave is varied continuously in accordance with the message. This is in direct contrast to pulse modulation, in which some parameter of a pulse train is varied in accordance with the message. There are two basic types of pulse modulation: pulse-analog modulation and pulse-code modulation. In pulse-analog modulation systems, a periodic pulse train is used as the carrier wave, and some characteristic feature of each pulse (e.g., amplitude, duration, or position) is varied in a continuous manner in accordance with the pertinent sample value of the message. On the other hand, in pulse-code modulation, a discrete-time, discrete-amplitude representation is used for the signal, and as such it has no cw counterpart.

A. Sampling Theorem

An operation that is basic to the design of all pulse-modulation systems is the sampling process whereby an analog signal is converted to a corresponding sequence of numbers that are usually uniformly spaced in time. For such a procedure to have practical utility, it is necessary that we choose the sampling rate properly, so that this sequence of numbers uniquely defines the original analog signal. This is the essence of the sampling theorem, which can be stated in two equivalent ways:

1. A band-limited signal of finite energy, which has no frequency components higher than W Hz, is completely described by specifying the values of the signal at instants of time separated by $1/2W$ s.
2. A band-limited signal of finite energy, which has no frequency components higher than W Hz, can be completely recovered from a knowledge of its samples taken at the rate of $2W$ per second.

The sampling rate of $2W$ samples per second, for a signal bandwidth of W Hz, is often called the Nyquist rate; its reciprocal $1/2W$ is called the Nyquist interval. The sampling theorem serves as the basis for the interchangeability of analog signals and digital sequences, which is so valuable in digital communication systems.

The derivation of the sampling theorem, as described above, is based on the assumption that the signal $g(t)$ is strictly band-limited. Such a requirement, however, can be satisfied only if $g(t)$ has infinite duration. In other words, a strictly band-limited signal cannot be simultaneously time-limited, and vice versa. Nevertheless, we can consider a time-limited signal to be essentially bandlimited in the sense that its frequency components outside some band of interest have negligible effects. We can then justify the practical application of the sampling theorem.

1. Aliasing

When the sampling rate $1/T_s$, exceeds the Nyquist rate $2W$, there is no problem in recovering the original signal $g(t)$ from its sampled version $g_\delta(t)$. When the sampling rate $1/T_s$ is less than $2W$, however, the aliasing effect arises. That is, a high-frequency component in the spectrum of the original signal seemingly takes on the identity of a lower frequency component in the spectrum of its sampled version. It is evident that if the sampling rate $1/T_s$ is less than the Nyquist rate $2W$, the original signal cannot be recovered exactly from its sampled version, and information is thereby lost in the sampling process.

Another factor that contributes to aliasing is the fact that a signal cannot be finite in both time and frequency, as mentioned previously. Since this violates the strict bandlimited requirement of the sampling theorem, we find that whenever a time-limited signal is sampled, there will always be some aliasing produced by the sampling process. Accordingly, in order to combat the effects of aliasing in practice, we use two corrective measures:

1. Before sampling, a low-pass anti-alias filter is used to attenuate those high-frequency components of the signal that lie outside the band of interest.
2. The filtered signal is sampled at a rate slightly higher than the Nyquist rate.

For example, a bandwidth of 3.4 kHz is considered adequate for the transmission of voice signals over a telephone network. In this basis, a sampling rate of 8 kHz (slightly

greater than twice 3.4 kHz) is considered the standard for the digital transmission of voice signals.

B. Pulse-Analog Modulation

In pulse-amplitude modulation (PAM), the amplitudes of regularly spaced rectangular pulses vary with the instantaneous sample values of a continuous message signal in a one-to-one fashion. This method of modulation is illustrated in Figs. 6a, b, and c, which represent a message signal, the pulse carrier, and the corresponding PAM wave, respectively.

The PAM wave $s(t)$ is easily demodulated by a low-pass filter with a cutoff frequency just large enough to accommodate the highest frequency component of the message signal $m(t)$. However, the reconstructed signal exhibits a slight amplitude distortion caused by the aperture effect due to lengthening of the samples.

In pulse-duration modulation (PDM), the samples of the message signal are used to vary the duration of the individual pulses. This form of modulation is also referred to as pulse-width modulation or pulse-length modulation. The modulating wave may vary with the time of occurrence of the leading edge, the trailing edge, or both edges of the pulse. In Fig. 6d the trailing edge of each pulse is varied in accordance with the message signal.

In PDM, long pulses expend considerable power during the pulse while bearing no additional information. If this unused power is subtracted from PDM, so that only time transitions are preserved, we obtain a more efficient type of pulse modulation known as pulse-position modulation (PPM). In PPM the position of a pulse relative to its unmodulated time of occurrence is varied in accordance with the message signal (Fig. 6e).

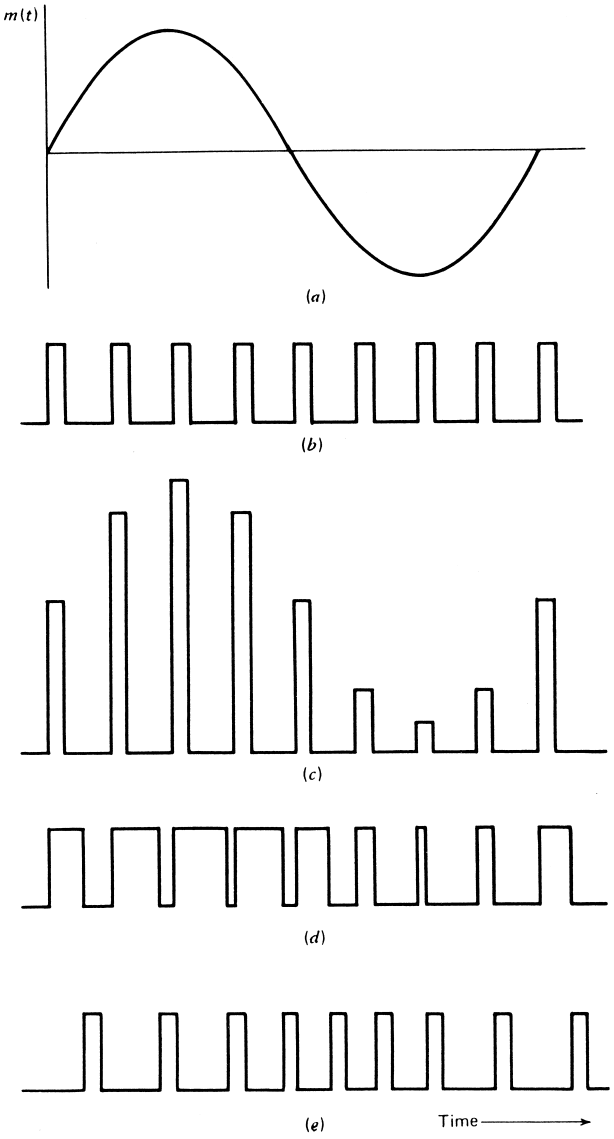


FIGURE 6 Different forms of pulse-analog modulation for the case of a sinusoidal modulating wave. (a) Modulating wave. (b) Pulse carrier. (c) PAM wave. (d) PDM wave. (e) PPM wave. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

C. Pulse-Code Modulation

In PAM, PDM, and PPM, only time is expressed in discrete form, whereas the respective modulation parameters (namely, pulse amplitude, duration, and position) are varied in a continuous manner in accordance with the message. Thus, in these modulation systems, information transmission is accomplished in analog form at discrete times. On the other hand, in pulse-code modulation (PCM), the message signal is sampled and the amplitude of each sample is rounded off to the nearest one of a finite set of allowable values, so that both time and amplitude are in discrete form. This allows the message to be transmitted by means of coded electrical signals, thereby distinguishing PCM from all other methods of modulation.

The use of digital representation of analog signals (e.g., voice, video) offers the following advantages: (1) ruggedness to transmission noise and interference, (2) efficient regeneration of the coded signal along the transmission path, and (3) the possibility of a uniform format for different kinds of baseband signals. These advantages, however, are attained at the cost of increased transmission bandwidth requirement and increased system complexity. With the increasing availability of wide-band communication channels, coupled with the emergence of the requisite device technology, the use of PCM has become a practical reality.

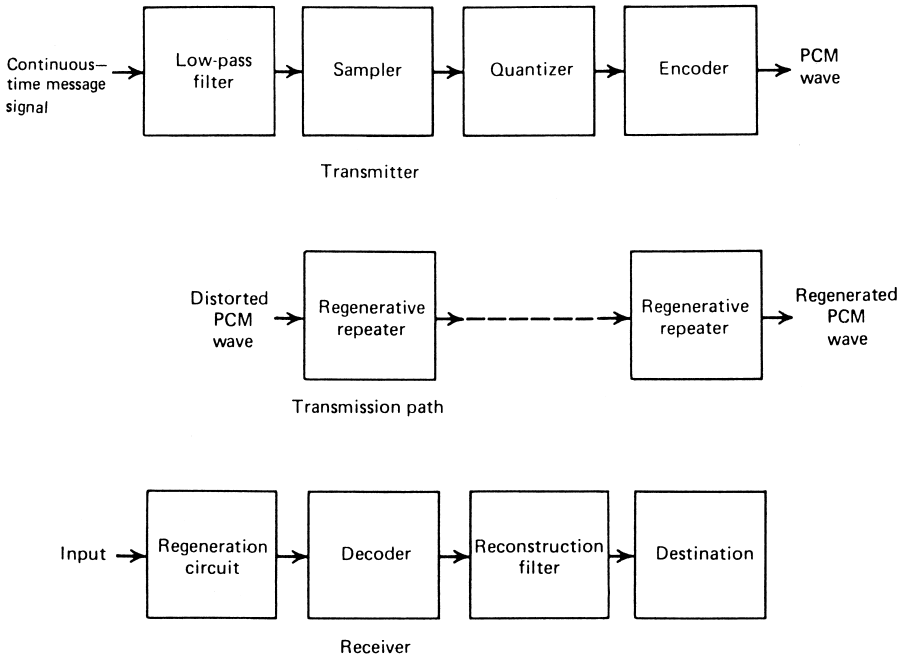


FIGURE 7 Basic elements of a PCM system. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

1. Elements of PCM

Pulse-code modulation systems are considerably more complex than PAM, PDM, and PPM systems, in that the message signal is subjected to a greater number of operations. The essential operations in the transmitter of a PCM system are sampling, quantizing, and encoding (Fig. 7). The quantizing and encoding operations are usually performed in the same circuit, which is called an analog-to-digital converter. The essential operations in the receiver are regeneration of impaired signals, decoding, and demodulation of the train of quantized samples. Regeneration usually occurs at intermediate points along the transmission route as necessary.

2. Sampling

The incoming message wave is sampled with a train of narrow rectangular pulses so as to approximate closely the instantaneous sampling process. In order to ensure perfect reconstruction of the message at the receiver, the sampling rate must be greater than twice the highest frequency component W of the message wave in accordance with the sampling theorem. In practice, a low-pass filter is used at the front end of the sampler in order to exclude frequencies greater than W before sampling. Thus, the application of sampling permits the reduction of the continuously varying message wave to a limited number of discrete values per second.

3. Quantizing

A continuous signal, such as voice, has a continuous range of amplitudes, and therefore its samples have a continuous amplitude range. In other words, within the finite amplitude range of the signal we find an infinite number of amplitude levels. It is not necessary in fact to transmit the exact amplitudes of the samples. Any human sense (the ear or the eye), as ultimate receiver, can detect only finite intensity differences. This means that the original continuous signal can be approximated by a signal constructed of discrete amplitudes selected on a minimum-error basis from an available set. The existence of a finite number of discrete amplitude levels is a basic condition of PCM. Clearly, if we assign the discrete amplitude levels with sufficiently close spacing, we can make the approximated signal practically indistinguishable from the original continuous signal.

The conversion of an analog (continuous) sample of the signal to a digital (discrete) form is called the quantizing process.

4. Encoding

In combining the processes of sampling and quantizing, the specification of a continuous baseband signal becomes limited to a discrete set of values, but not in the form best suited to transmission over a line or radio path. To exploit the advantages of sampling and quantizing, we require the

use of an encoding process to translate the discrete set of sample values to a more appropriate form of signal. Any plan for representing each of this discrete set of values as a particular arrangement of discrete events is called a code. One of the discrete events in a code is called a code element or symbol. For example, the presence or absence of a pulse is a symbol. A particular arrangement of symbols used in a code to represent a single value of the discrete set is called a code word or character.

In a binary code, each symbol can be either of two distinct values or kinds, such as the presence or absence of a pulse. The two symbols of a binary code are customarily denoted 0 and 1. In a ternary code, each symbol can be one of three distinct values or kinds, and so on for other codes. However, the maximum advantage over the effects of noise in a transmission medium is obtained by using a binary code, because a binary symbol withstands a relatively high level of noise and is easy to regenerate.

Suppose that in a binary code, each code word consists of n bits (*bit* is an acronym for binary digit). Then, using such a code, we can represent a total of 2^n distinct numbers. For example, a sample quantized into one of 128 levels can be represented by a seven-bit code word.

5. Regeneration

The most important feature of PCM systems is the ability to control the effects of distortion and noise produced by transmitting a PCM wave through a channel. This capability is accomplished by reconstructing the PCM wave by means of a chain of regenerative repeaters located at sufficiently close spacing along the transmission route.

6. Decoding

The first operation in the receiver is to regenerate (i.e., reshape and clean up) the received pulses one last time. These clean pulses are then regrouped into code words and decoded (i.e., mapped back) into a quantized PAM signal. The decoding process involves generating a pulse the amplitude of which is the linear sum of all the pulses in the code word, with each pulse weighted by its place value (2^0, 2^1, 2^2, 2^3, ...) in the code.

7. Filtering

The final operation in the receiver is to recover the signal wave by passing the decoder output through a low-pass reconstruction filter whose cutoff frequency is equal to the message bandwidth W. Assuming that the transmission path is error free, the recovered signal includes no noise, with the exception of the initial distortion introduced by the quantization process.

8. Noise in PCM Systems

The performance of a PCM system is influenced by two major sources of noise:

1. Transmission noise, which may be introduced anywhere between the transmitter output and the receiver input. The effect of transmission noise is to introduce bit errors into the received PCM wave, with the result that in the case of a binary system, a symbol 1 occasionally is mistaken for a symbol 0, or vice versa. Clearly, the more frequently such errors occur, the more dissimilar to the original message signal is the receiver output. The fidelity of information transmission by PCM in the presence of transmission noise is conveniently measured in terms of the error rate, or probability of error, defined as the probability that the symbol at the receiver output differs from that transmitted.
2. Quantizing noise, which is introduced in the transmitter and is carried along to the receiver output. Typically, each bit in the code word of a PCM system contributes 6 dB to the signal-to-noise ratio at the receiver output.

Unlike frequency modulation, PCM offers an exponential law for the trade-off between transmission bandwidth and noise performance. It is for this reason that PCM is the preferred method for the transmission of voice and video signals.

D. Differential Pulse-Code Modulation

When a voice or video signal is sampled at a rate higher than the Nyquist rate, the resulting sampled signal exhibits high correlation between adjacent samples. The meaning of this high correlation is that in an average sense, the signal does not change rapidly from one sample to the next, with the result that the difference between adjacent samples has a variance that is smaller than the variance of the signal itself. When these highly correlated samples are encoded, as in a standard PCM system, the resulting encoded signal contains redundant information. This means that symbols that are not absolutely essential to the transmission of information are generated as a result of the encoding process. By removing this redundancy before encoding, we obtain a more efficient coded signal.

Now, if we know a sufficient part of a redundant signal, we can infer the rest, or at least make the most probable estimate. In particular, if we know the past behavior of a signal up to a certain point in time, it is possible to make some inference about its future values; such a process is commonly called prediction. The fact that it is possible to predict future values of the signal $m(t)$ provides motivation for the differential quantization scheme shown in Fig. 8. The input to the quantizer is a signal that consists of the

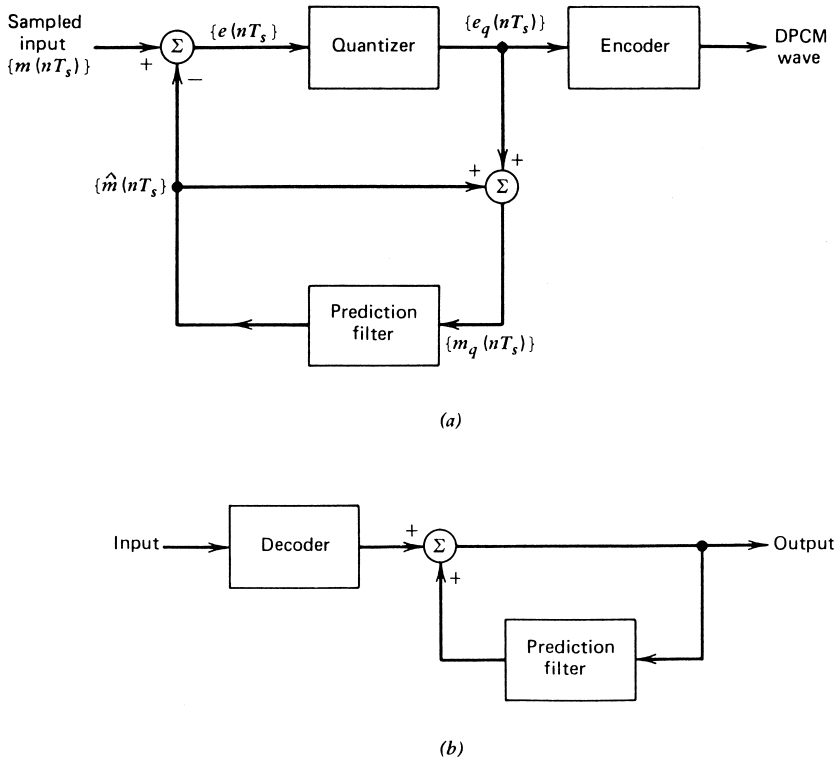


FIGURE 8 DPCM system. (a) Transmitter. (b) Receiver. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John wiley & Sons, Inc.]

difference between the unquantized input symbol $m(nT_s)$. This predicted value is produced by using a prediction filter. The difference signal $e(nT_s)$ is called a prediction error. By encoding the quantizer output, as in Fig. 8a, we obtain a variation of PCM known as differential pulse-code modulation (DPCM). This encoded signal is used for transmission.

The receiver for reconstructing the quantized version of the input is shown in Fig. 8b. It consists of a decoder to reconstruct the quantized error signal. The quantized version of the original input is reconstructed from the decoder output using the same design of prediction filter in the transmitter.

E. Delta Modulation

The exploitation of signal correlations in DPCM suggests the further possibility of oversampling a baseband signal (i.e., at a rate much higher than the Nyquist rate) purposely to increase the correlation between adjacent samples of the signal, so as to permit the use of a simple quantizing strategy for constructing the encoded signal. Delta modulation (DM), which is the one-bit (or two-level) version of DPCM, is precisely such a scheme.

In its simple form, DM provides a staircase approximation to the oversampled version of an input baseband signal (Fig. 9). The difference between the input and the approximation is quantized into only two levels, namely, $\pm\delta$, corresponding to positive and negative differences, respectively. Thus, if the approximation falls below the signal at any sampling epoch, it is increased by δ. If, on the other hand, the approximation lies above the signal, it is diminished by δ. Provided that the signal does not change too rapidly from sample to sample, the staircase approximation remains within $\pm\delta$ of the input signal.

The principal virtue of DM is its simplicity. It can be generated by applying the sampled version of the incoming baseband signal to a modulator that involves a summer, quantizer, and accumulator interconnected (Fig. 10).

In comparing the DPCM and DM networks of Figs. 8 and 10, respectively, we note that they are basically similar, except for two important differences, namely, the use of a one-bit (two-level) quantizer in the delta modulator and the replacement of the prediction filter by a single delay element.

IV. DATA TRANSMISSION

In this section we consider some of the issues involved in transmitting digital data (of whatever origin) over a communication channel. The transmission may be in baseband

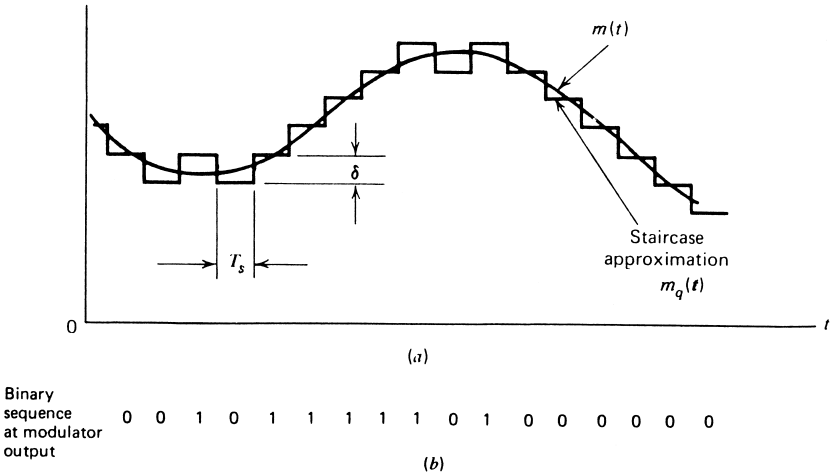


FIGURE 9 Delta modulation. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

or passband form. For baseband transmission, we use discrete pulse modulation, in which the amplitude, duration, or position of the transmitted pulses is varied in a discrete manner in accordance with the digital data. However, it is customary to use discrete PAM because it is the most efficient in terms of power and bandwidth utilization.

To transmit digital data over bandpass channels (e.g., satellite channels) we require the use of digital modulation techniques whereby the amplitude, phase, or frequency of a sinusoidal carrier is varied in a discrete manner in accordance with the digital data. In practice, the variation of phase or frequency of the carrier is the preferred method.

A. Baseband Data Transmission

The basic elements of a baseband binary PAM system are shown in Fig. 11. The signal applied to the input of the

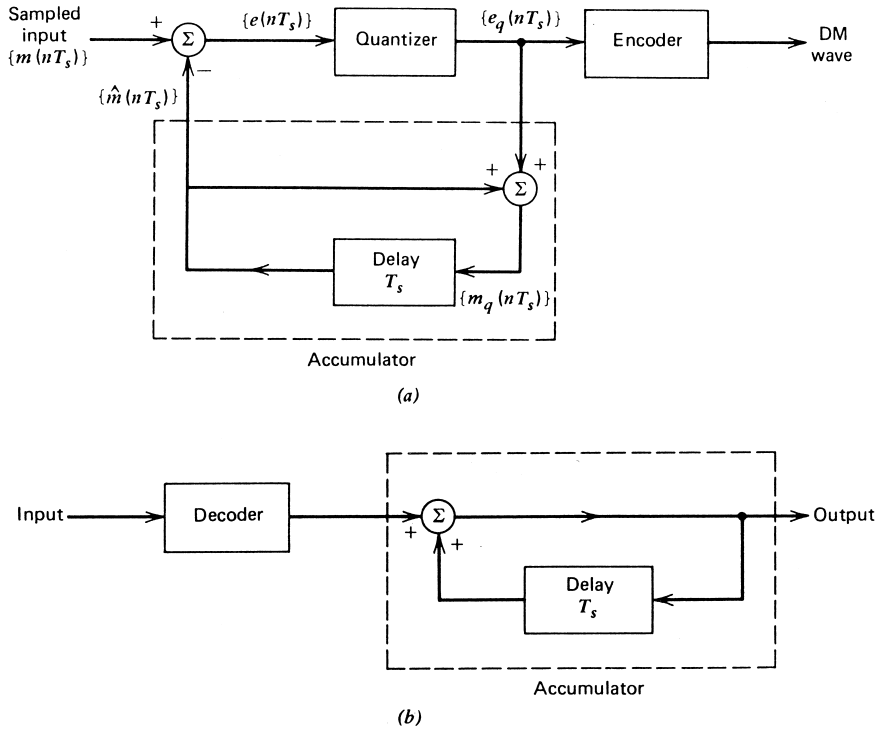


FIGURE 10 DM system. (a) Transmitter. (b) Receiver. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

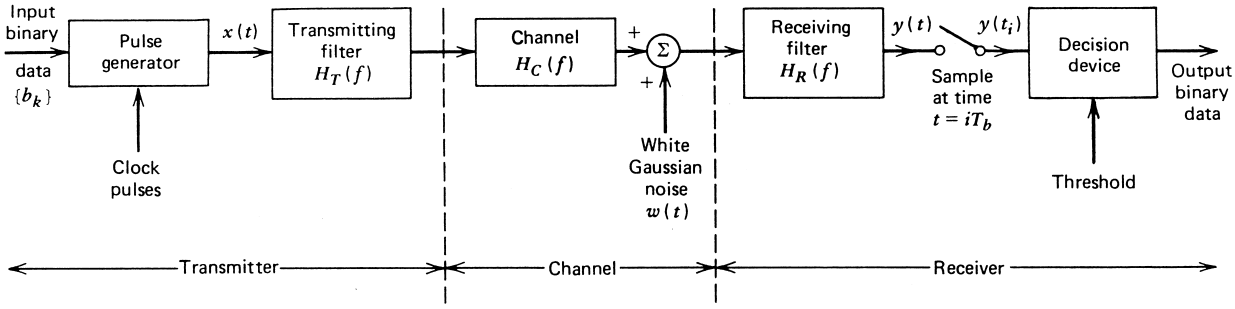


FIGURE 11 Baseband binary data transmission system. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

system consists of a binary data sequence $[b_k]$ with a bit duration of T_b seconds; b_k is in the form of 1 or 0. This signal is applied to a pulse generator, producing the pulse waveform

$$x(t) = \sum_{k=-\infty}^{\infty} a_k g(t - kT_b) \tag{14}$$

where $g(t)$ denotes the shaping pulse that is normalized, so that we write

$$g(0) = 1 \tag{15}$$

The amplitude a depends on the identity of the input bit b; specifically, we assume that

$$a_k = \begin{cases} +a & \text{if the input bit } b_k \text{ is represented by symbol 1} \\ -a & \text{if the input bit } b_k \text{ is represented by symbol 0} \end{cases} \tag{16}$$

The PAM signal $x(t)$ passes through a transmitting filter of transfer function $H_T(f)$. The resulting filter output defines the transmitted signal, which is modified in a deterministic fashion as a result of transmission through the channel of transfer function $H_C(f)$. In addition, the channel adds random noise to the signal at the receiver input. Then the noisy signal is passed through a receiving filter of transfer function $H_R(f)$. This filter output is sampled synchronously with the transmitter, the sampling instants being determined by a clock or timing signal that is usually extracted from the receiving filter output. Finally, the sequence of samples thus obtained is used to reconstruct the original data sequence by means of a decision device. The amplitude of each sample is compared with a threshold. If the threshold is exceeded, a decision is made in favor of symbol 1 (say). If the threshold is not exceeded, a decision is made in favor of symbol 0. If the sample amplitude equals the threshold exactly, the flip of a fair coin will determine which symbol was transmitted.

1. Baseband Shaping

Typically, the transfer function of the channel and the pulse shape are specified, and the problem is to determine the transfer functions of the transmitting and receiving filters so as to enable the receiver to recognize the sequence of values in the received signal wave.

In solving this problem, we have to overcome the intersymbol interference (ISI) caused by the overlapping tails of other pulses adding to the particular pulse $A_i p(t - iT_b)$, which is examined at the sampling time iT_b. If this form of interference is too strong, it may result in erroneous decisions in the receiver. Clearly, control of ISI in the system is achieved in the time domain by controlling the function $p(t)$.

The receiving filter output $y(t)$ is sampled at time $t = iT_b$ (with i taking on integer values), yielding

$$\begin{aligned} y(t_i) &= \sum_{k=-\infty}^{\infty} A_k p[(i-k)T_b)] + n(t_i) \\ &= A_i + \sum_{\substack{k=-\infty \\ k \neq i}}^{\infty} A_k p[(i-k)T_b)] + n(t_i) \end{aligned} \tag{17}$$

where $A_k p(t)$ is the response of the cascade connection of the transmitting filter, the channel, and the receiving filter, which is produced by the pulse $a_k g(t)$ applied to the input of the cascade. In Eq. (17), the first term A_i represents the ith transmitted bit. The second term represents the residual effect of all other transmitted bits on the decoding of the ith bit; this residual effect is called the intersymbol interference. The last term $n(t_i)$ represents the noise sample at time t_i.

In the absence of ISI and noise, we observe from Eq. (17) that

$$y(t_i) = A_i$$

which shows that under these conditions, the ith transmitted bit can be decoded correctly. The unavoidable presence

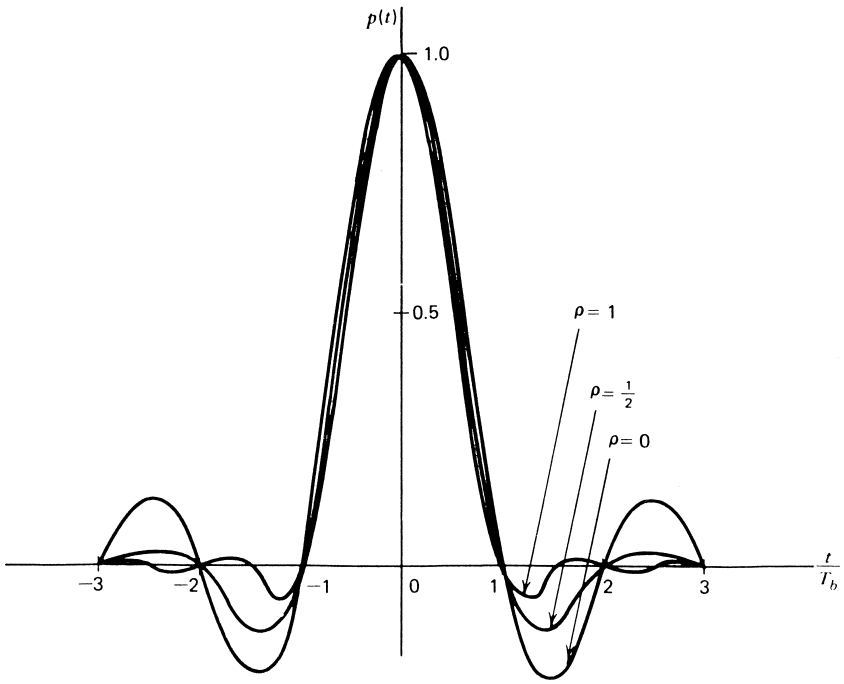


FIGURE 12 Time response $p(t)$ for varying rolloff factor ρ. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

of ISI and noise in the system, however, introduces errors in the decision device at the receiver output. Therefore, in the design of the transmitting and receiving filters, the objective is to minimize the effects of noise and ISI and thereby deliver the digital data to their destination in an error-free manner as far as possible.

One signal waveform that produces zero ISI in a realizable fashion is defined by

$$p(t) = \text{sinc}(2B_\text{T}t)\frac{\cos(2\pi\rho B_\text{T}t)}{1-16\rho^2B_T^2t^2} \tag{18}$$

where $\text{sinc}(2B_\text{T}t)$ is the sinc function

$$\text{sinc}(2B_\text{T}t) = \frac{\sin(2\pi B_\text{T}t)}{2\pi B_\text{T}t} \tag{19}$$

and the transmission bandwidth B_T is related to the bit duration T_b as

$$B_\text{T} = 1/2T_\text{b}$$

The parameter ρ is called the rolloff factor; its value lies within the range $0 \le p \le 1$.

The function $p(t)$ of Eq. (18) consists of the product of two factors: the factor $\text{sinc}(2B_\text{T}t)$ associated with an ideal low-pass filter and a second factor that decreases as $1/|t|^2$ for large $|t|$. The first factor ensures zero crossings of $p(t)$ at the desired sampling instants of time $t = iT$, with i an integer (positive and negative). The second factor reduces the tails of the pulse considerably below that obtained from the ideal low-pass filter, so that the transmission of binary waves using such pulses is relatively insensitive to sampling time errors. In fact, the amount of ISI resulting from this timing error decreases as the rolloff factor ρ is increased from zero to unity.

The time response $p(t)$ is plotted in Fig. 12 for $\rho = 0$, 0.5, and 1. For the special case of $\rho = 1$, the function $p(t)$ simplifies as

$$p(t) = \frac{\text{sinc}(4B_\text{T}t)}{1-16B_T^2t^2} \tag{20}$$

This time response exhibits two interesting properties:

1. At $t = \pm T_\text{b}/2 = \pm 1/4B_\text{T}$, we have $p(t) = 0.5$; that is, the pulse width measured at half-amplitude is exactly equal to the bit duration T_b.
2. There are zero crossings at $t = \pm 3T_\text{b}/2$, $\pm 5T_\text{b}/2$, ... in addition to the usual zero crossings at the sampling times $t = \pm T_\text{b}, \pm 2\text{T}_\text{b}, \ldots$.

These two properties are particularly useful in generating a timing signal from the received signal for the purpose of synchronization. However, this requires the use of a transmission bandwidth double that required for the ideal case corresponding to $\rho = 0$.

2. Correlative Coding

Thus far we have treated ISI as an undesirable phenomenon that produces a degradation in system performance.

Indeed, its very name connotes a nuisance effect. Nevertheless, by adding ISI to the transmitted signal in a controlled manner, it is possible to achieve a signaling rate of $2B_T$ symbols per second in a channel of bandwidth of B_T hertz. Such schemes are called correlative coding or partial-response signaling schemes. The design of these schemes is based on the premise that since ISI introduced into the transmitted signal is known, its effect can be interpreted at the receiver. Thus, correlative coding can be regarded as a practical means of achieving the theoretical maximum signaling rate of $2B_T$ symbols per second in a bandwidth of B hertz, as postulated by Nyquist, using realizable and perturbation-tolerant filters.

3. Adaptive Equalization

An efficient approach to the high-speed transmission of digital data (e.g., computer data) over a voice-grade telephone channel (which is characterized by a limited bandwidth and high signal-to-noise ratio) involves the use of two basic signal processing operations:

1. Discrete PAM by encoding the amplitudes of successive pulses in a periodic pulse train with a discrete set of possible amplitude levels.
2. A linear modulation scheme that offers bandwidth conservation (e.g., VSB) to transmit the encoded pulse train over the telephone channel.

At the receiving end of the system, the received wave is demodulated and then synchronously sampled and quantized. As a result of dispersion of the pulse shape by the channel, the number of detectable amplitude levels is often limited by ISI rather than by additive noise. In principle, if the channel is known precisely, it is virtually always possible to make the ISI (at the sampling instants) arbitrarily small by using a suitable pair of transmitting and receiving filters, so as to control the overall pulse shape in the manner described above. The transmitting filter is placed directly before the modulator, whereas the receiving filter is placed directly after the demodulator. Thus, insofar as ISI is concerned, we can consider the data transmission to be essentially baseband.

However, in a switched telephone network, two factors contribute to the distribution of pulse distortion on different link connections: (1) differences in the transmission characteristics of the individual links that can be switched together and (2) differences in the number of links in a connection. The result is that the telephone channel is random in the sense of being one of an ensemble of possible channels. Consequently, the use of a fixed pair of transmitting and receiving filters designed on the basis of average channel characteristics may not adequately reduce ISI. To realize the full transmission capability of a telephone channel, there is need for adaptive equalization. By *equalization* we mean the process of correcting channel-in-duced distortion. This process is said to be adaptive when it adjusts itself continuously during data transmission by operating on the input signal.

Among the philosophies for adaptive equalization of data transmission systems are prechannel equalization at the transmitter and postchannel equalization at the receiver. Because the first approach requires a feedback channel, it is customary to use only adaptive equalization at the receiving end of the system. This equalization can be achieved, before data transmission, by training the filter with the guidance of a suitable training sequence transmitted through the channel so as to adjust the filter parameters to optimum values. The typical telephone channel changes little during an average data call, so that precall equalization with a training sequence is sufficient in most cases encountered in practice. The equalizer is positioned after the receiving filter in the receiver.

4. Eye Pattern

One way to study ISI in a PCM or data transmission system experimentally is to apply the received wave to the vertical deflection plates of an oscilloscope and to apply a sawtooth wave at the transmitted symbol rate $1/T$ to the horizontal deflection plates. The waveforms in successive symbol intervals are thereby translated into one interval on the oscilloscope display. The resulting display is called an eye pattern because, for binary waves, its appearance resembles the human eye. The interior region of the eye pattern is called the eye opening.

An eye pattern provides a great deal of information about the performance of the pertinent system (Fig. 13):

1. The width of the eye opening defines the time interval over which the received wave can be sampled without

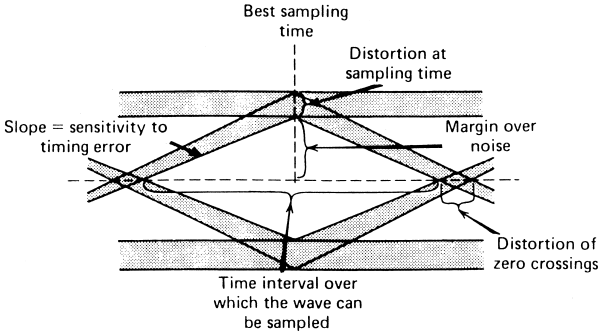


FIGURE 13 Interpretation of the eye pattern. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

error from ISI. It is apparent that the preferred time for sampling is the instant of time at which the eye is open widest.

2. The sensitivity of the system to timing error is determined by the rate of closure of the eye as the sampling time is varied.

3. The height of the eye opening, at a specified sampling time, defines the margin over noise.

When the effect of ISI is severe, traces from the upper portion of the eye pattern cross traces from the lower portion, with the result that the eye is completely closed. In such a situation, it is impossible to avoid errors due to the combined presence of ISI and noise in the system.

5. Digital Subscriber Lines

A digital subscriber line (DSL) operates over a local loop (less than 1.5 km) that provides a direct connection between a user terminal (e.g., computer) and a telephone company's central office. In this way, a user is connected to a broadband backbone data network, which is based on technologies such as the asynchronous transfer mode (ATM) and Internet Protocol (IP). The information-bearing signal is thereby kept in the digital domain all the way from the user terminal to an Internet service provider, with the signal being switched or routed at regular intervals in the course of its transmission through the data network.

A major source of channel impairment in DSLs is the near-end crosstalk (NEXT), which is generated by transmitters at the same end of the cable as the receiver. The other major source of impairment is ISI that arises due to the imperfect frequency response of the channel.

A widely used line code for DSLs is the 2BIQ codes, which stands for a pair of binary digits encoded into one quaternary symbol. This code is a block code representing a four-level PAM signal.

Asymmetric digital subscriber lines (ADSL) are designed to simultaneously support three services on a single twisted-wire pair:

1. Data transmission downstream (toward the subscriber)
2. Data transmission upstream (away from the subscriber)
3. Plain old telephone service

The DSL is said to be "asymmetric" because the downstream bit rate is much higher than the upstream bit rate.

B. Passband Data Transmission

When digital data are to be transmitted over a bandpass channel, it is necessary to modulate the incoming data onto a carrier wave (usually sinusoidal) with fixed frequency limits imposed by the channel. The data may represent digital computer outputs or PCM waves generated by digitizing voice or video signals, and so on. The channel may be a microwave radio link or satellite channel, and so on. In any event, the modulation process involves switching or keying the amplitude, frequency, or phase of the carrier in accordance with the incoming data. Thus, there are three basic signaling techniques: amplitude-shift keying (ASK), frequency-shift keying (FSK), and phase-shift keying (PSK), which can be viewed as special cases of amplitude modulation, frequency modulation, and phase modulation, respectively. Figure 14 illustrates these three basic signaling techniques for binary data input.

Ideally, FSK and PSK signals have a constant envelope. This feature makes them impervious to amplitude nonlinearities, as encountered in microwave radio links and satellite channels. Accordingly, in practice, FSK and PSK signals are much more widely used than ASK signals.

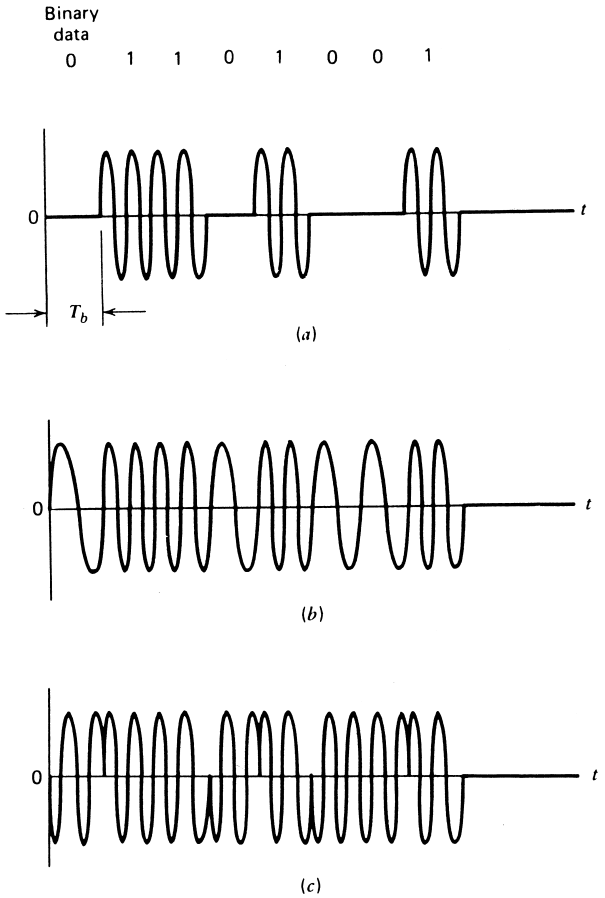


FIGURE 14 Three basic forms of signaling binary information. (a) Amplitude-shift keying. (b) Frequency-shift keying. (c) Phase-shift keying. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

The system is said to be coherent when the locally generated carrier in the receiver is synchronized to that in the transmitter in both frequency and phase. It is said to be noncoherent when the phase of the incoming signal is destroyed in the receiver.

1. Coherent Binary PSK

In a coherent binary PSK system, a pair of signals, $s_1(t)$ and $s_2(t)$, are used to represent binary symbols 1 and 0, respectively; they are defined by

$$s_1(t) = \sqrt{2E_b/T_b}\cos(2\pi f_c t) \tag{21a}$$

$$s_2(t) = \sqrt{2E_b/T_b}\cos(2\pi f_c t + \pi) = -\sqrt{2E_b/T_b}\cos(2\pi f_c t) \tag{21b}$$

where $0 \le t \le T_b$ and E_b is the transmitted signal energy per bit. In order to ensure that each transmitted bit contains an integral number of cycles of the carrier wave, the carrier frequency f_c is chosen equal to n_c/T_b for some fixed integer n_c. A pair of sinusoidal waves differing only in phase by 180°, as defined above, are referred to as antipodal signals.

2. Coherent Binary FSK

In a binary FSK system, the symbols 1 and 0 are distinguished from one another by the transmission of one of two sinusoidal waves that differ in frequency by a fixed amount. A typical pair of sinusoidal waves is described by

$$s_i(t) = \begin{cases} \sqrt{2E_b/T_b}\cos(2\pi f_i t), & 0 \le t \le T_b \\ 0, & \text{elsewhere} \end{cases} \tag{22}$$

where the carrier frequency f_i equals one of two possible values: f_1 and f_2. The transmission of frequency f_1 represents symbol 1, and the transmission of frequency f_2 represents symbol 0.

3. Coherent Quadrature Signaling Techniques

Channel bandwidth and transmitted power constitute two primary "communication resources," the efficient utilization of which provides the motivation for the search for spectrally efficient modulation schemes. The primary objective of spectrally efficient modulation is to maximize the bandwidth efficiency, defined as the ratio of data rate to channel bandwidth (in units of bits per second per hertz) for a specified probability of symbol error. A secondary objective is to achieve this bandwidth efficiency at a minimum practical expenditure of average signal power or, equivalently, in a channel perturbed by additive white Gaussian noise, a minimum practical expenditure of average signal-to-noise ratio.

Examples of the quadrature-carrier multiplexing system include the following signaling techniques:

1. Quadriphase-shift keying (QPSK), which is an extension of the binary PSK. In this technique, the phase of the carrier takes on one of four possible values. Specifically, the binary pairs 10, 00, 01, 11 are assigned the phase values $\pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$, respectively. This is illustrated in Fig. 15.
2. Minimum shift keying (MSK), which is a special form of continuous-phase frequency-shift keying, with the detection in the receiver being performed in two successive bit intervals. The nominal carrier frequency f_c is equal to the arithmetic mean of the two frequencies f_1 and f_2 used to represent symbols 1 and 0, respectively. Moreover, the carrier frequency f_c is chosen equal to an integral multiple of one-fourth of the bit rate in order to make the phase of the transmitted signal continuous at the bit transition instants. One other requirement is to make the deviation ratio

$$h = T_b(f_1 - f_2)$$

equal to one-half. Figure 16 illustrates the wave-form of the MSK signal.

4. *M*-Ary Signaling Techniques

In an M-ary signaling scheme, we send any one of M possible signals, $s_1(t), s_2(t), \ldots, s_M(t)$, during each signaling interval of duration T. For almost all applications, the number of possible signals M is 2^n, where n is an integer. The symbol duration T is equal to nT_b, where T_b is the bit duration. These signals are generated by changing

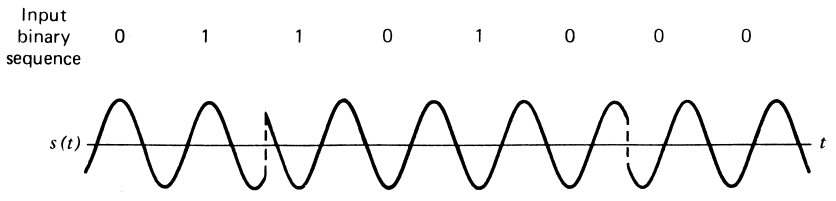


FIGURE 15 QPSK wave $s(t)$. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

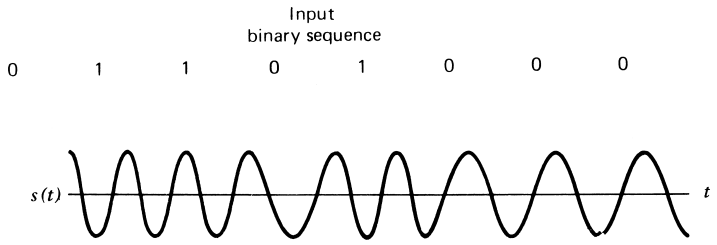


FIGURE 16 MSK wave $s(t)$. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

the aplitude, phase, or frequency of a carrier in M discrete steps. Thus, we have M-ary ASK, M-ary PSK, and M-ary FSK digital modulation schemes. The QPSK system considered above is an example of M-ary PSK with $M = 4$.

M-Ary signaling schemes are preferred over binary signaling schemes for transmitting digital information over bandpass channels when the requirement is to conserve bandwidth at the expense of increasing power. In practice, a communication channel rarely has the exact bandwidth required for transmitting the output of an information source by means of binary signaling schemes. Thus, when the bandwidth of the channel is less than the required value, we use M-ary signaling schemes so as to utilize the channel efficiently.

5. Detection of Signals with Random Phase in the Presence of Noise

Up to this point, we have assumed that the information-bearing signal is completely known at the receiver. In practice, however, in addition to the uncertainty due to the additive noise of a receiver, there is often an additional uncertainty due to the randomness of certain signal parameters. The usual cause of this uncertainty is distortion in the transmission medium. Perhaps the most common random signal parameter is the phase, which is especially true for narrow-band signals.

For example, transmission over a multiplicity of paths of different, and variable lengths or rapidly varying delays in the propagating medium from transmitter to receiver may cause the phase of the received signal to change in a way that the receiver cannot follow. Synchronization with the phase of the transmitted carrier may then be too costly, and the designer may simply choose to disregard the phase information in the received signal at the expense of some degradation in the noise performance of the system.

For example, in the noncoherent detection of binary FSK signals, one can simplify receiver complexity by using a pair of frequency discriminators tuned to the frequencies f_1 and f_2, which represent binary symbols 0 and 1, respectively.

Another example is that of differential phase-shift keying (DPSK), which can be viewed as the noncoherent version of binary PSK. It eliminates the need for a coherent reference signal at the receiver by combining two basic operations at the transmitter: (1) differential encoding of the input binary wave and (2) phase-shift keying—hence the name differential phase-shift keying. In effect, to send symbol 0 we phase-advance the current signal waveform by 180°, and to send symbol 1 we leave the phase of the current signal waveform unchanged. The receiver is equipped with a storage capability, so that it can measure the relative phase difference between the waveforms received during two successive bit intervals. Provided that the unknown phase contained in the received wave varies slowly (i.e., slowly enough to be considered essentially constant over two bit intervals), the phase difference between waveforms received in two successive bit intervals will be independent of θ.

6. Noise Comparison of Digital Modulation Schemes Using a Single Carrier

It is customary to use the average probability of symbol error to assess the noise performance of digital modulation schemes. It should be realized, however, that even if two systems yield the same average probability of symbol error, their performances, from the user's viewpoint, may be quite different. In particular, the greater the number of bits per symbol, the more the bit errors will cluster together. For example, if the average probability of symbol error is 10^{-3}, the expected number of symbols occurring between any two erroneous symbols is 1000. If each symbol represents one bit of information (as in a binary PSK or binary FSK system), the expected number of bits separating two erroneous bits is 1000. If, on the other hand, there are two bits per symbol (as in a QPSK system), the expected separation is 2000 bits. Of course, a symbol error generally creates more bit errors in the second case, so that the percentage of bit errors tends to be the same. Nevertheless, this clustering effect may make one system more attractive than another, even at the same symbol error rate. In

the final analysis, which system is preferable will depend on the situation.

Two systems having an unequal number of symbols can be compared in a meaningful way only if they use the same amount of energy to transmit each bit of information. It is the total amount of energy needed to transmit the complete message, not the amount of energy needed to transmit a particular symbol satisfactorily, that represents the cost of the transmission. Accordingly, in comparing the different data transmission systems considered, we shall use as the basis of our comparison the average probability of symbol error expressed as a function of the bit energy-to-noise density ratio E_b/N_0.

In Table I, we have summarized the expressions for the average probability of symbol error P_e for the coherent PSK, conventional coherent FSK with one-bit decoding, DPSK, noncoherent FSK, QPSK, and MSK, when operating over an AWGN channel. In Fig. 17 we have used these expressions to plot P_e as a function of E_b/N_0.

In summary, we can state the following:

1. The error rates for all the systems decrease monotonically with increasing values of E_b/N_0.
2. For any value of E_b/N_0, the coherent PSK produces a smaller error rate than any of the other systems. Indeed, in the case of systems restricted to one-bit decoding perturbed by additive white Gaussian noise, the coherent PSK system is the optimum system for transmitting binary data in the sense that it achieves the minimum probability of symbol error for a given value of E_b/N_0.
3. The coherent PSK and the DPSK require an E_b/N_0 that is 3 dB less than the corresponding values for the conventional coherent FSK and the noncoherent FSK, respectively, to realize the same error rate.
4. At high values of E_b/N_0, the DPSK and the noncoherent FSK perform almost as well (to within ~1 dB) as the coherent PSK and the conventional coherent FSK, respectively, for the same bit rate and signal energy per bit.
5. In QPSK two orthogonal carriers $\sqrt{2/T}\cos(2\pi f_c t)$ and $\sqrt{2/T}\sin(2\pi f_c t)$ are used, where the carrier frequency f_c is an integral multiple of the symbol rate $1/T$, with the result that two independent bit streams can be transmitted and subsequently detected in the receiver. At high values of E_b/N_0, coherently detected binary PSK and QPSK have the same error rate performances for the same value of E_b/N_0.
6. In MSK the two orthogonal carriers $2\sqrt{T_b}\cos(2\pi f_c t)$ and $2\sqrt{T_b}\sin(2\pi f_c t)$ are modulated by antipodal symbol shaping pulses $\cos(\pi t/2T_b)$ and $\sin(\pi t/2T_b)$, respectively, over $2T_b$ intervals, where T_b is the bit duration. Correspondingly, the receiver uses a coherent phase decoding process over two successive bit intervals to recover the original bit stream. Thus, MSK has exactly the same error rate performance as QPSK.

TABLE I Summary of Formulas for the Symbol Error Probability for Different Data Transmission Systems

System	Error probability P_e
Coherent binary signaling	
Coherent PSK	$\frac{1}{2}\operatorname{erfc}(\sqrt{E_b/N_0})$
Coherent FSK	$\frac{1}{2}\operatorname{erfc}(\sqrt{E_b/2N_0})$
Noncoherent binary signaling	
DPSK	$\frac{1}{2}\exp(-E_b/N_0)$
Noncoherent FSK	$\frac{1}{2}\exp(-E_b/2N_0)$
Coherent quadrature signaling	
QPSK	$\operatorname{erfc}(\sqrt{E_b/N_0})$
MSK	$-\frac{1}{4}\operatorname{erfc}^2(\sqrt{E_b/N_0})$

7. Discrete Multitone

The ADSL, described in Section IV.A.5, is a data transmission system capable of realizing megabit per second rates over existing twisted-pair telephone lines. An elegant modulation technique well suited for this application is discrete multitone (DMT), which allows the modulator characteristics to be a function of measured channel characteristics. The basic idea of DMT is rooted in a commonly used engineering principle: divide and conquer, whereby a difficult problem is solved by dividing it into a number of simpler problems and then combining the solutions.

In DMT, a wideband channel is transformed into a set of N subchannels that operate in parallel. What makes DMT distinctive is the fact that the transformation is performed in discrete time as well as discrete frequency. Consequently, the input–output behavior of the entire communication system admits a linear matrix representation, which lends itself to implementation using the discrete Fourier transform.

Figure 18 shows a block diagram of the DMT data-transmission system. The transmitter consists of the following functional blocks:

- Demultiplexer, which converts the incoming serial data stream into parallel form.
- Constellation encoder, which maps the parallel data into $N/2$ multibit subchannel with each subchannel being represented by a specific signal constellation (e.g., M-ary quadrature amplitude modulation).
- Inverse discrete Fourier transform (IDFT), which transforms the frequency-domain parallel data at the constellation encoder output into parallel time-domain

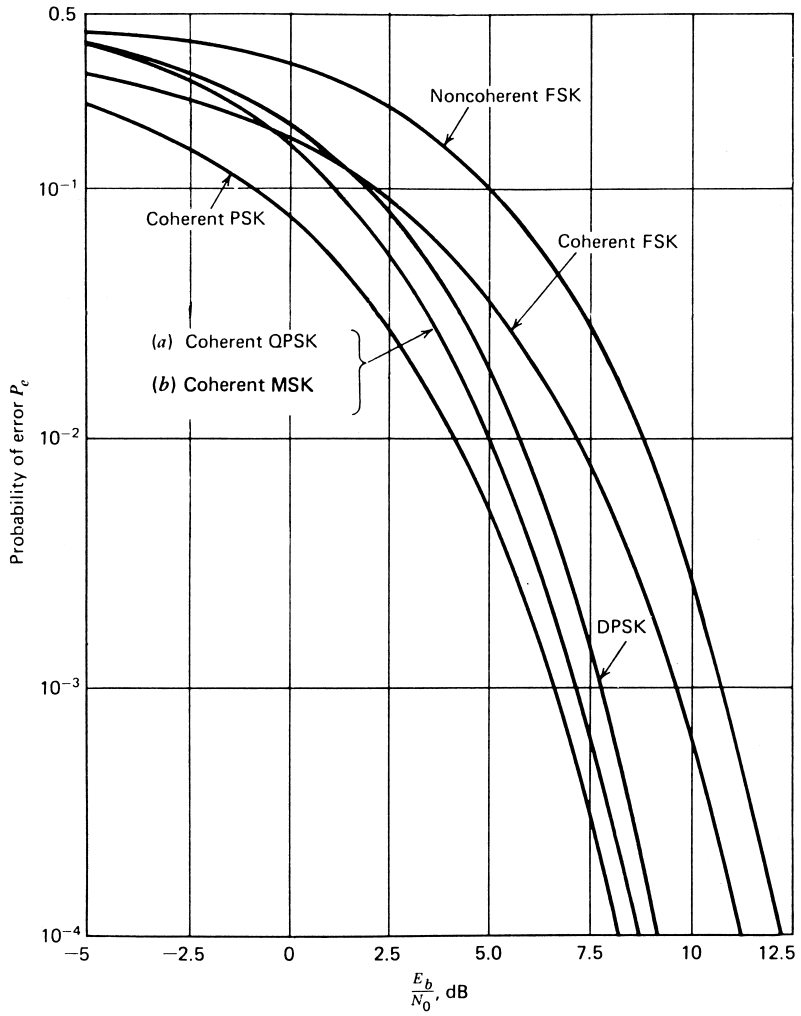


FIGURE 17 Comparison of the noise performances of different PSK and FSK systems. [From Haykin, S. (1983). "Communication Systems," 2nd ed. Wiley, New York. © John Wiley & Sons, Inc.]

data. For efficient implementation of the IDFT, it is customary to use the fast Fourier transform algorithm by choosing $N = 2^k$ where k is a positive integer.

- Digital-to-analog converter (DAC), which converts the digital data into an analog form for transmission over the channel.

The receiver consists of an analog-to-digital converter (DAC), serial-to-parallel converter, discrete Fourier transformer (DFT), decoder, and multiplexer, which perform the inverse operations of the transmitter. A reconstruction of the original data stream is achieved.

An important feature of DMT is that it permits the realization of a capacity of data transmission in accordance with Shannon's information capacity theorem:

$$C = \frac{1}{2}\log_2(1 + \text{SNR})$$

where the capacity C is measured in bits per transmission and SNR is the (dimensionless) received signal-to-noise ratio at the channel output. This remarkable performance is realized through the use of loading, which refers to the process of allocating transmit power to the N individual subchannels so as to maximize the bit rate of the entire multichannel transmission system.

8. Orthogonal Frequency-Division Multiplexing

DMT is one particular form of multichannel modulation. Another closely related form of multichannel modulation is orthogonal frequency-division multiplexing (OFDM), which differs from DMT in areas of application and some aspects of its design.

OFDM is used for data transmission over radio broadcast channels and wireless communication channels. This domain of application requires some changes to the design

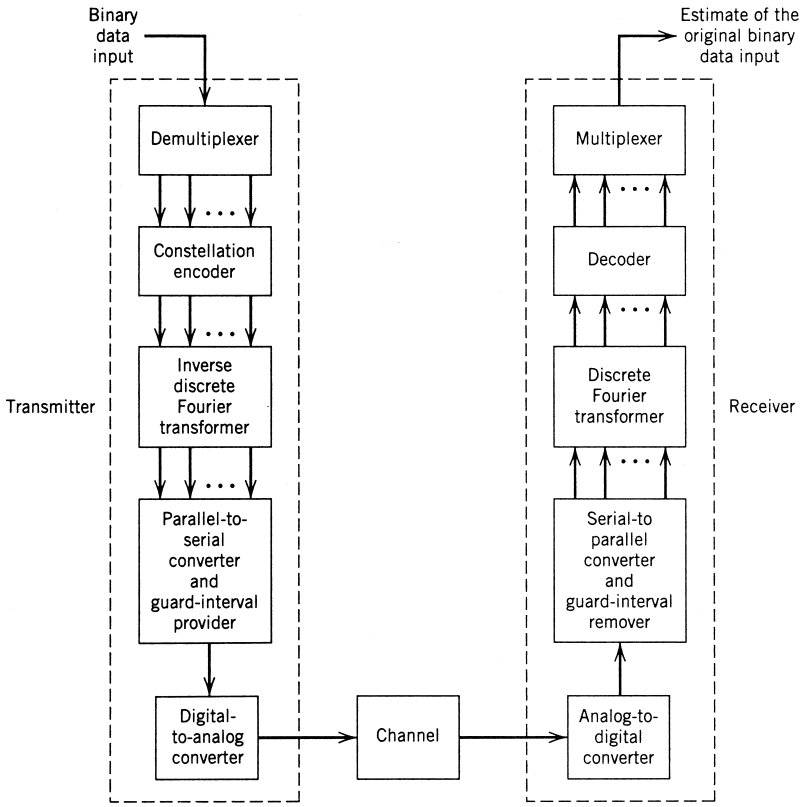


FIGURE 18 Block diagram of the discrete-multitone (DMT) data-transmission system.

of the OFDM system. Unlike DMT, which uses loading for bit allocation, OFDM uses a fixed number of bits per subchannel. This restriction is made necessary by the fact that a broadcast channel involves one-way transmission, and in a wireless communication environment the channel is varying too rapidly.

The other modification in OFDM involves the use of (1) an upconverter after the ADC in Fig. 18 to translate the transmitted frequency to facilitate signal propagation over a radio wireless channel, and (2) a downconverter before the ADC to undo the frequency translation performed by the upconverter in the transmitter.

V. SPREAD-SPECTRUM MODULATION

A major issue of concern in the study of digital communications is that of providing for the efficient utilization of bandwidth and power. Notwithstanding the importance of these two primary communication resources, there are situations where it is necessary to sacrifice this efficiency in order to meet certain other design objectives. For example, the system may be required to reject interference as in wireless communications. This requirement is catered to by a class of signaling techniques known as spread-spectrum modulation.

In the so-called direct-sequence spread-spectrum technique, two stages of modulation are used. First, the incoming data sequence is used to modulate a wide-band binary code, which transforms this narrow-band data sequence into a noiselike wide-band signal. The resulting wide-band signal undergoes a second modulation using a phase-shift keying technique (e.g., QPSK). For its operation, such a spread-spectrum modulation system relies on the availability of a noiselike spreading code called a pseudonoise sequence.

At the receiver, the channel output passes through two stages of demodulation: QPSK detection, followed by despreading. The net result is a processing gain (PG) due to spreading, which is defined by

$$\text{PG} = \frac{T_b}{T_c}$$

where T_b is the bit duration of the data sequence at the transmitter input, and T_c is the chip duration, that is, the duration of binary symbol 1 or 0 in the spreading code.

VI. MULTIPLEXING

Multiplexing is the process of combining several message signals (originating from independent users) for their

transmission over a common channel. Three commonly used methods of multiplexing are as follows:

1. Frequency-division multiplexing (FDM), in which CW modulation (e.g., frequency modulation) is used to translate each message signal to reside inside a specific frequency slot inside the passband of the channel by assigning it a distinct carrier frequency. At the receiver, a bank of filters is used to separate the different modulated signals and prepare them individually for demodulation.
2. Time-division multiplexing (TDM), in which some form of pulse modulation (e.g., pulse code modulation) is used to position samples of the different message signals in nonoverlapping time slots.
3. Code-division multiplexing (CDM), in which each message signal is identified by a distinctive spreading code.

Thus, in FDM the message signals appear across the channel all the time. In contrast, in TDM each message signal occupies the full passband of the channel. CDM differs from both FDM and TDM in that each message signal occupies the full passband of the channel all the time; for demultiplexing to function satisfactorily, the spreading codes assigned to the individual message signals should ideally be orthogonal to each other.

ACKNOWLEDGMENT

The material presented in this article is summarized from the author's book, "Communication Systems" (see Bibliography). The author is indebted to John Wiley and Sons, publisher of this book, for permission to write this article.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SATELLITE SYSTEMS • DATA TRANSMISSION MEDIA • NETWORKS FOR DATA COMMUNICATION • RADIO SPECTRUM UTILIZATION • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL

BIBLIOGRAPHY

Bellamy, J. C. (1991). "Digital Telephony," 2nd ed., Wiley, New York.
Gitlin, R. D., Hayes, J. F., and Weinstein, S. B. (1992). "Data Communication Principles," Plenum, New York.
Haykin, S. (2001). "Communication Systems," 4th ed. Wiley, New York.
Jayant, N. S., and Noll, P. (1984). "Digital Coding of Waveforms." Prentice-Hall, Englewood Cliffs, NJ.
Jeruchim, M. C., Balaban, B., and Shanmugan, J. S. (2000). "Simulation of Communication Systems," 2nd ed. Kluwer, Dordrecht, The Netherlands.
Lee, E. A., and Messerschmitt, D. G., (1994). "Digital Communication," 2nd ed. Kluwer Academic, Dordrechat, The Netherlands.
Proakis, J. G. (1995). "Digital Communication," 3rd ed., McGraw-Hill, New York.
Schwartz, M. (1980). "Information Transmission, Modulation and Noise," 3rd ed., McGraw-Hill, New York.
Starr, T., Cioffi, J. M., and Silverman, P. J. (1999). "Understanding Digital Subscriber Line Technology," Prentice-Hall, Englewood Cliffs; NJ.
Viterbi, A. J., and Omura, J. K. (1979). "Principles of Digital Communication and Coding," McGraw-Hill, New York.



Data Transmission Media

John S. Sobolewski
University of New Mexico

I. Basic Signal Types
II. Characteristics of Some Common Signals
III. Communication Channels
IV. Channel Capacity
V. Baseband Transmission
VI. Broadband Transmission
VII. Multiplexing
VIII. Types of Data Transmission Media
IX. Transmission Impairments
X. Bounded Transmission Media
XI. Unbounded Transmission
XII. Interconnection Considerations for Electrical Conductors
XIII. Standards
XIV. Summary

GLOSSARY

Bit Short for binary digit, the smallest unit of information in a binary system. A bit represents the choice between a one (mark) or a zero (space) condition.

Carrier Continuous frequency or periodic pulses capable of being modulated or changed by an information signal.

Channel Path for transmission of information between two or more points using some form of transmission medium.

Channel capacity Number of information bits per second that can be transmitted over a channel.

Demodulation Process of retrieving information from a modulated carrier wave. The opposite of modulation.

Medium, bounded Transmission medium that constrains and guides or conducts information-carrying signals, also called a guided transmission medium. Examples include wires and optical fibers.

Medium, transmission Transmission medium that provides the physical channel or path used for transmission of information between two or more points.

Medium, unbounded Transmission medium that permits signals to be transmitted but does not guide or constrain them; also called an unguided transmission

medium. Examples include radiowave transmission using space or air as the transmission medium.

Modulation Process of varying some characteristic of a carrier wave in accordance with the instantaneous value or sample of the information to be transmitted.

Multiplexing Division of a higher bandwidth communication channel into two or more lower bandwidth channels.

Repeater Device whose function is to retime and restore received signals back to their original form and strength and to retransmit them further down the channel.

Signal Aggregate of waves propagated along a transmission channel toward a receiver at the destination.

ELECTRONIC COMMUNICATION systems depend upon transmission media for sending voice, data, video, or other signals from one point to another. Among the physical media are twisted pairs, coaxial cable, free space, and optical fibers. The distance of transmission can be very small, as between computer memory and the central processing unit, or very large, as between a corporate headquarters and an overseas subsidiary. Without the transmission media, communication systems could not exist.

In recent years, there has been a virtual explosion in consumer demand for bandwidth due to the traffic generated by the Internet (300% growth per year), electronic mail, faxes, cellular telephones, mobile computers, teleconferencing, as well as conventional data and video transmission. This demand for new services and higher bandwidth has resulted in the deployment of new communication infrastructure and new technologies for the more effective use of the transmission media already in place. Every indication is that these trends will continue in the foreseeable future.

I. BASIC SIGNAL TYPES

Information is transmitted over communication channels either as analog signals or digital signals. Analog signals are in the form of a continuous time-varying physical quantity, such as voltage magnitude or frequency, that reflects the variations of the information or signal source with time. Speech, radio, and conventional television signals are examples of analog signals. Digital signals consist of a stream of discrete pulses with well-defined states. Data generated by digital computers and their peripherals or data terminals are digital in nature and, in general, consist of two discrete levels corresponding to digital ones and zeros, as illustrated in Fig. 1.

It is important to note that a communication system can be designed to carry either analog or digital signals by using appropriate signal conversion devices. This applies to systems using any type of transmission medium—wire, coaxial cable, optical fibers, or free space. Furthermore, since analog signals can be converted to digital signals and vice versa, any type of information can be transmitted either in digital or analog form that is compatible with the transmission medium over which it will travel with no significant loss of information.

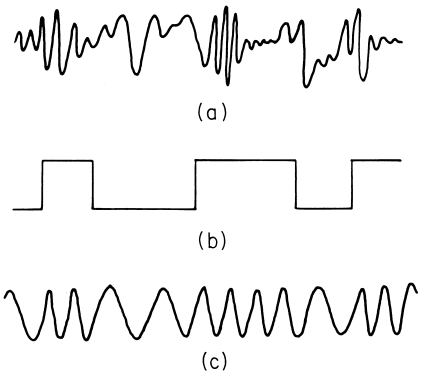


FIGURE 1 Any information can be transmitted in (a) analog or (b) digital form. Conversion from one to the other is available; (c) shows a digital signal converted to an analog signal using frequency modulation.

In the past, most communication channels were designed for transmission of analog signals within a certain range of frequencies. If we need to transmit digital computer data for long distances over such channels, we can use a special device that converts the digital data into an analog signal consisting of a continuous range of frequencies, as illustrated in Fig. 1. We can thus use existing analog channels for transmission of digital data. Conversely, it is possible to convert analog voice or television signals into digital signals and transmit them over channels designed for digital signals. The capability of converting analog to digital signals and digital to analog signals is important, since the need for transmission of digital data is growing rapidly while many of the communication channels in existence are analog in nature. However, digital transmission has advantages over analog transmission, and consequently digital technology is evolving rapidly. Over the past two decades, millions of miles of digital transmission channels have been installed by communication carriers, mostly to carry data and digitized speech. This trend will continue, and communication channels installed in the future will be almost exclusively digital in nature, not only for transmission of data but also because digital channels are more cost effective for virtually all communication needs.

II. CHARACTERISTICS OF SOME COMMON SIGNALS

Before describing the concept of a communication channel and the various transmission media in greater detail,

it is necessary to describe the characteristics of some of the most common types of signals that are routinely transmitted over existing communication channels. Even though significant economic payoff can result from designing communication channels that can carry all types of signals, some of the signal characteristics are so different that different communication architectures are needed. The sections that follow describe the characteristics of program and speech, which are analog signals; data, which is a digital signal; and video, which can be considered a hybrid between analog and digital.

A. Program Signals

Program signals include radio broadcasts, the audio portion of television programs, and high-fidelity music programs that are distributed over cable systems to subscribing customers. Such signals are characterized by a greater average volume, dynamic range in volume, duration, and bandwidth (which results in greater fidelity) than typical telephone speech signals. A high dynamic range of volume is needed to reproduce faithfully the variety of program material transmitted, which may include speech, music, and special sound effects. The bandwidth required may be as high as 30 to 15,000 Hz for high-quality music for FM stereo or the audio portion of television programs, even though only a fraction of this bandwidth is used for speech during, say, a newscast.

B. Speech Signals

Speech signals are usually transmitted over telephone channels. The telephone set transforms the voice signals into electrical analog signals which are transmitted to the local telephone exchange and then through the wide-area telephone network to the receiving party. Although speech typically covers frequencies from 30 to 10,000 Hz, most of the energy is in the range from 200 to 3500 Hz. Since the human ear is not very sensitive to small changes in frequencies and since humans can correct for things such as missing syllables or words, we do not need to reproduce speech signals precisely to achieve acceptable quality of transmission. Consequently, most telephone communications are bandlimited to between 200 and 3500 Hz to save transmission costs. This savings comes about since costs are directly related to the bandwidth (i.e., the range of frequencies transmitted) and (as will be described later) bandlimiting allows a greater number of voice channels to be multiplexed over a high-bandwidth channel. It must be pointed out, however, that the nominal bandwidth of a voice channel is defined as 4000 Hz, with the additional bandwidth allowing a guard band on either side of the speech signal to reduce interference between channels.

The time required to set up a telephone connection can vary from a few seconds to tens of seconds, the communication is almost always two-way, and the two parties continuously transmit (talk), listen (receive), or pause until the call is terminated. A wide dynamic range in volume is needed, although not as large as for program signals. Conversations require immediate delivery of the signal (i.e., delays are not tolerated), but communication is relatively tolerant of noise on the channel.

C. Data Signals

Servers, workstations, or personal computers usually produce digital signals consisting of streams of binary digits (bits) representing ones or zeros. Unlike the speech signal, which contains much redundant information and is relatively tolerant of noise or errors on the channel, the digital data signal may have little or no redundancy; therefore, errors cannot be tolerated. In fact, bits (parity or cyclic redundancy bits) are usually added to the basic digital signal to detect errors and correct them by retransmission. By addition of even more bits and use of error-correcting codes, errors can be detected and corrected without retransmission of the data.

Other important characteristics are shown in Table I, which summarizes the differences between digital and

TABLE I Differences in Characteristics between Data and Telephone Voice Signals

Data signal	Telephone voice signal
Desirable to set up a connection in one second or less	One second to tens of seconds to set up a connection
One- or two-way transmission	Two-way transmission in most cases
Error-free received data	Tolerant of noise and some errors on the channel
Little or no redundancy	Much inherently redundant information
Transmission usually in bursts	Transmit or receive continuously until call is disconnected
Data can usually be stored and transmitted later when convenient	Not tolerant of transmission delays
Transmission has high peaks, peak-to-average ratio as high as 1000 to 1	Transmission rate relatively constant
Connection may be required 24 hours/day, 7 days/week (e.g., bank cash machine)	Duration of connection usually several minutes
May require a wide range of bandwidths, from thousands to tens of millions of bits per second	Requires fixed bandwidth of about 4000 Hz

telephone voice signals. It is important to recognize that those differences are significant and that telephone and computer users have fundamentally different requirements. With the large increase in data transmission since the beginning of the computer era, and the even greater increases projected for the future, there will be continued rapid progress in the development of new network architectures designed specifically for data transmission.

D. Video Signals

Most video signals carried over communication channels today are color television signals. A video transmission system must deal with four important factors when transmitting images of moving objects, specifically:

1. Perception of the distribution of luminance (degree of shade between light and dark)
2. Perception of a three-dimensional image (width, height, and depth)
3. Perception of motion related to the two factors above
4. Perception of color (hues and tints)

Monochrome video transmission deals with the first three factors. In color video transmission, color is combined with the monochrome (black and white) picture. At every moment, the video signal must integrate luminance and color from a scene in three dimensions, as a function of time, and combine them into a complex electrical signal for transmission. What makes this more complex is that time itself is variable since the scene is changing all the time. The integration of visual information is carried out by a process called *scanning*.

The scanning process consists of taking a horizontal strip across the image and scanning it from left to right beginning at the top. When the right-hand end of a strip is reached, the next lower horizontal strip is scanned. Luminance values are translated on each scanning interval into voltage or current variations and are transmitted. In the U.S., a total of 525 horizontal scan lines make a whole picture frame. Scanning a whole frame is repeated at a rate of 30 frames per second. This rate takes advantage of the persistence of vision of the human eye, thus eliminating flicker and giving the perception of motion, as in motion pictures. Because of the amount of information that must be transmitted to produce flicker-free images with acceptable horizontal and vertical resolution, video signals require a bandwidth of 6 MHz in the U.S., as compared with a 4-kHz bandwidth required for acceptable transmission of voice signals.

To permit decoding of the video signal at the receiver, it is necessary to transmit vertical and horizontal synchronizing pulses interleaved with the picture information. These pulses synchronize the transmitter and receiver by identifying the start of every horizontal scan line that corresponds to the top of the picture and the subsequent horizontal scan lines that make up the rest of the picture. Since these synchronizing pulses are essentially digital signals while the picture information in each horizontal scan line is in analog form, video signals may be thought of as hybrid signals.

III. COMMUNICATION CHANNELS

A block diagram of an electronic communication system is shown in Fig. 2. It consists of an input signal, a transmitter, a communication channel, a receiver, and an output signal from the receiver. The transmitter modifies the input message signal into a form suitable for transmission over the channel, which is the transmission path for providing communication between transmitter and receiver. The purpose of the receiver is to recreate the original message signal at the output. For example, when digital data signals must be sent over a communication channel designed primarily for analog signals, the transmitter has to convert the digital signals to analog signals by a process known as *modulation*. The receiver then demodulates the analog signal back to digital form and passes it on to its ultimate destination.

The communication channel can consist of various media such as wire, coaxial cable, optical fibers, or free space, in which case the signal is radiated as an electromagnetic wave as in conventional television or radio broadcasting.

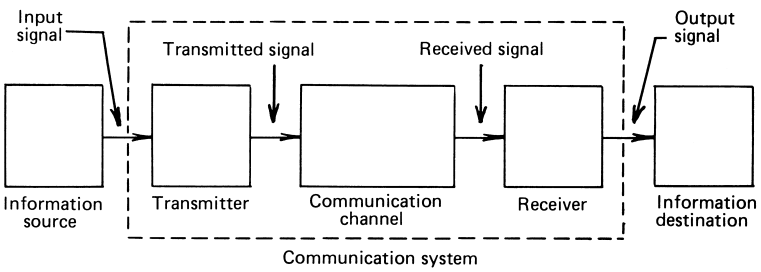


FIGURE 2 Model of a communication system.

Communication channels have practical limitations, such as bandwidth, and suffer from various impairments, such as nonlinearities as well as noise and interference, which may be introduced from within or from outside the channel. Transmitters and receivers must be carefully designed to match the signals to be transmitted to the physical properties of the communication channel and to minimize the effects of channel impairments on the quality of reception.

We can classify communication channels as analog or digital depending on the basic signals they transmit. Analog channels can be further classified as voice, program, or video channels depending on their intended use and the kind of signals they carry. They may be characterized by their bandwidth, which is the range of frequencies they transmit. Thus, a voiceband channel may be called a 4-kHz channel while broadband channels may be 48- or 240-kHz channels. Digital channels are usually characterized by their bit rate. Channels with a bit rate of 1540 kbps (kilobits per second), for example, are used with increasing frequency and can be used to transmit simultaneously several lower bit rate channels through multiplexing, as explained in Section VII.

Analog and digital channels can also be classified as simplex, half-duplex, or full-duplex. Simplex channels transmit in one direction only and are used for radio broadcast but seldom for data communication, which usually requires two-way transmission. Half-duplex channels can send in both directions but only in one direction at any given time. In other words, they provide nonsimultaneous two-way communication. Full-duplex channels, on the other hand, allow simultaneous two-way communication.

IV. CHANNEL CAPACITY

In practice, a communication channel represents a financial investment, hence the goal of communication engineers is to maximize the return on that investment. In the case of digital transmission, this is done by maximizing the channel capacity, which may be defined as the maximum rate at which information can be sent over the channel with an arbitrarily small probability of error.

In 1928, Nyquist showed that for binary transmission (i.e., the transmitted signal has one of two possible values at any one time) $2W$ bits/sec can be transmitted over a channel of bandwidth W in the absence of noise. In the general case, if we use M-ary ($M > 2$) rather than pure binary transmission by sending one of M different and distinguishable signal values at any one time, then in the absence of noise we have

$$C = 2W \log_2 M \tag{1}$$

where C is the channel capacity in bits/sec and W the channel bandwidth in Hz. Thus, if we have a channel bandwidth W of 3300 Hz, then in the absence of noise we can transmit up to 6600 bits/sec using pure binary transmission ($M = 2$). However, if we use a transmission scheme that makes it possible to send one of four distinguishable values at any given instant ($M = 4$), then we increase the channel capacity to 13,200 bits/sec. Similarly, if we can send one of eight distinguishable values at any instant, then each level in effect represents $\log_2 8 = 3$ bits, and the channel capacity becomes 19,800 bits/sec using this transmission scheme.

For a channel of bandwidth W, it is highly desirable to increase M in the above equation to maximize the channel capacity. The question arises, however, as to the number of different signaling values that can be transmitted and be separately distinguished at the receiver in the presence of noise, distortion, limits on signal power, and other channel impairments that occur in practice. In 1948, Shannon proved that if signals are sent with signal power S over a channel with Gaussian noise (amplitude of noise signal follows a Gaussian distribution) of power N, then the capacity C of the channel in bits per second is

$$C = W \log_2(1 + S/N) \tag{2}$$

where W is the bandwidth of the channel. This is one of the fundamental laws of communication and gives the maximum signaling rate over a communication channel in terms of three parameters that are known or measurable. We can design elaborate coding schemes or modulation techniques, but we will never be able to increase the capacity unless we increase either the available bandwidth or the signal-to-noise ratio S/N.

If we take the previous example of the channel with a bandwidth of 3300 Hz and if we assume a signal-to-noise ratio of 63, then the maximum possible rate at which data could be transmitted over this channel is $3300 \log_2(1 + 63) = 19{,}800$ bits/sec, no matter how many different signaling levels our transmission scheme may have at any one time. In fact, for this particular channel, if we tried to use a transmission scheme with one of the 16 distinguishable signal values at any instant to increase the channel capacity, then in the absence of noise, relation (1) gives a capacity of 26,400 bits/sec. However, with the given signal-to-noise ratio of 63, the capacity is limited to the 19,800 bits/sec, implying that 16 different values would not be distinguishable on this channel unless the signal-to-noise ratio was increased, which could be done by increasing the signal power of the transmitted signal.

Unfortunately, as explained in Section VIII, practical communication channels include impairments other than Gaussian noise, and the limit given by relation (2) is very difficult or costly to achieve in practice. However, it clearly identifies the two resources that designers can use to increase the channel capacity, namely, channel bandwidth and transmitted signal power. Sometimes these resources

are limited, in which case the channels may be classified as bandlimited, as in the case of telephone voice channels, or power-limited, as in the case of, say, a satellite channel.

V. BASEBAND TRANSMISSION

Signals can be transmitted over a communication channel using either baseband or broadband transmission. In baseband transmission, the information signal is sent over the channel directly without modification. In broadband transmission, the information signal is modified by superimposing it on a higher-frequency signal, called the *carrier*, which "carries" the information over the channel. Broadband transmission, or carrier transmission as it is sometimes called, uses modulation techniques that are described in the next section.

Because there is no modification of the information signal, baseband transmission systems can be used for digital or analog signals. In practice, they require an electrical or light (e.g., optical fiber) conductor to carry the baseband signal and are used primarily for transmission over relatively short distances. Separate transmitters and receivers are seldom required (e.g., the source and destination act as transmitter and receiver, respectively), and in those cases in which they are required they are relatively simple devices. The simplicity of baseband transmission makes it very common. In fact, it is used for short-distance signal transmission amongst most components of electronic devices or systems. Transmission of signals between components within a computer, a computer and its local peripheral devices, and an audio amplifier and the speakers are all examples of baseband transmission. Others include transmission over local area networks in a building or factory and over telephone subscriber loops between a telephone set and the local exchange. Longer telephone links, such as those between exchanges, do not usually carry baseband signals but use higher carrier frequencies to carry many signals simultaneously.

Because of the relatively short distances involved, the bandwidth of baseband channels can be relatively high. The major limitations of such channels include stray capacitance, inductance, and resistance which cause distortion, especially of the leading and trailing edges of digital signals. These limitations will be described in greater detail in Sections IX and XII.

VI. BROADBAND TRANSMISSION

In broadband transmission, information signals are processed and superimposed on a carrier that is more suitable for transmission over a particular channel. This is known as *modulation*, which is defined as the process by which some characteristic (e.g., amplitude, frequency, or phase) of a carrier wave is varied or modulated in accordance with the instantaneous value or samples of the information signal to be transmitted. The information signal, which can be analog or digital, is called the *modulating wave*, while the result of the modulation process is called the *modulated wave*. Demodulation restores the modulated signal to its original form at the receiving end of the communication channel.

At first, modulation techniques were used primarily for radio broadcasts where a relatively low-bandwidth speech or music program signal was used to modulate a much higher frequency carrier, resulting in a modulated signal whose frequency spectrum was suited for radio transmission. Subsequently these techniques were used for telephone transmission, because it was realized that telephone lines had a bandwidth greater than that needed for speech. By using several voice signals to modulate carriers of different frequencies, several voice signals can be transmitted (or multiplexed) over a single telephone line, thus reducing the cost per voice channel.

Modulation is used for two other important reasons. First, it is used to convert digital signals for transmission over analog channels that would otherwise destroy the digital signal. Second, it is used to convert voice or other analog signals to digital form to achieve lower noise or less distortion. Modulation, detection, and demodulation are complex topics, and only the most basic principles can be described here. A more complete treatment is given in Schwartz (1980) and Proakis (1983).

A. Analog Modulation

In analog modulation, a sine-wave carrier is modulated by an analog signal such as voice of a program signal. The sine-wave carrier $c(t)$ may be represented by:

$$c(t) = A_c \sin(2\pi f_c t + \theta_c) \tag{3}$$

where A_c is the carrier amplitude, f_c is the carrier frequency, and θ_c is the phase. The value of A_c, f_c, and θ_c can be varied by the message or modulating signal $m(t)$ to form the modulated wave $s(t)$. In all cases, the carrier frequency f_c must be much greater than the highest frequency component of the modulating signal $m(t)$.

In amplitude modulation (AM), the amplitude of the carrier is varied in accordance with the message signal $m(t)$. This is done by adding the product of $c(t)$ and a fraction k of $m(t)$ to the carrier $c(t)$, yielding:

$$\begin{aligned} s(t) &= c(t) + km(t)c(t) \\ &= A_c(1 + km(t)) \sin(2\pi f_c + \theta_c) \end{aligned} \tag{4}$$

The absolute value of $s(t)$ has the same shape as $m(t)$ provided that the absolute value of $km(t)$ is always less than 1.

By taking the Fourier transform of $s(t)$, it may be shown that it consists of two symmetrical sidebands, each of bandwidth m, on either side of the carrier frequency f_c, where m is the bandwidth of the message signal $m(t)$. Consequently, one of the sidebands is sometimes removed through filtering to conserve bandwidth, and the carrier is suppressed to reduce transmission power. This variation of AM is known as *single sideband suppressed carrier* (SSB/SC) modulation and was commonly used for voice and data transmission over telephone circuits.

In frequency modulation (FM), the amplitude of $s(t)$ is constant, but the instantaneous frequency $f_i(t)$ of the modulated wave $s(t)$ changes linearly with the amplitude of the message signal $m(t)$ and is given by:

$$f_i(t) = f_c + k_f m(t) \tag{5}$$

which yields

$$s(t) = A_c \sin(2\pi(f_c + k_f m(t))t + \theta_c) \tag{6}$$

where k_f is a constant and determines the maximum frequency deviation that can occur. In phase modulation (PM), the amplitude and frequency are maintained constant, but the instantaneous phase of the modulated wave $s(t)$ changes linearly with $m(t)$, yielding:

$$s(t) = A_c \sin(2\pi f_c t + k_\theta m(t)) \tag{7}$$

where k_θ is a constant and determines the maximum phase shift that can occur.

Amplitude and frequency modulation are used widely for radio program signals. An important feature of frequency and phase modulation is that they are more immune to noise and interference than amplitude modulation. This advantage, however, is achieved at the expense of a more complex receiver and greater bandwidth requirement. In fact, frequency and phase modulation provide a convenient means for obtaining better noise performance at the expense of bandwidth. Figure 3 illustrates the three modulation techniques in the case where the modulating signal is a digital rather than an analog signal.

B. Digital–Analog Modulation

The three analog modulation techniques just described can also be used for digital signals. In the case of binary digital signals, the modulation process involves keying (switching) the amplitude, frequency, or phase of the carrier between two different values according to the message signal. This is a form of digital–analog modulation and results in three distinct techniques known as *amplitude shift keying* (ASK), *frequency shift keying* (FSK), and *phase shift keying* (PSK). They are illustrated in Fig. 3 and are extensively used in older modems (modulator–demodulators) which allow digital signals to be transmitted over analog channels. Figure 4 illustrates a simple demodulation (or detection) scheme for amplitude shift keying. The

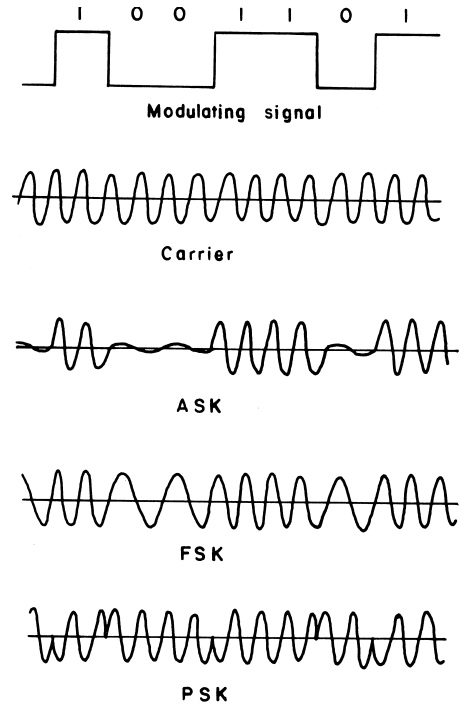


FIGURE 3 Three basic techniques of modulating a sinusoidal carrier using a digital modulating signal: ASK, FSK, and PSK.

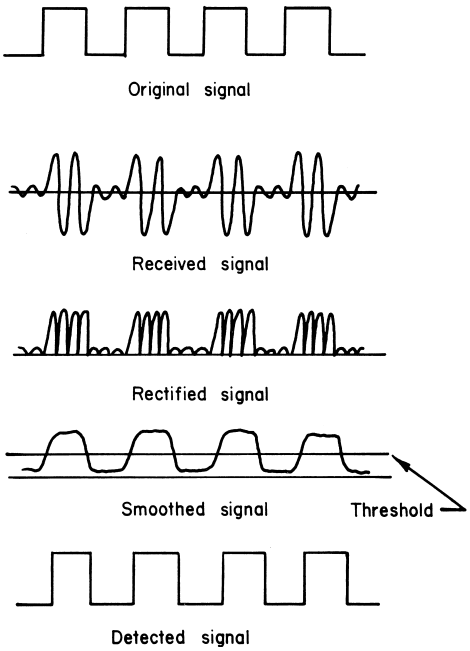


FIGURE 4 Demodulation used with ASK.

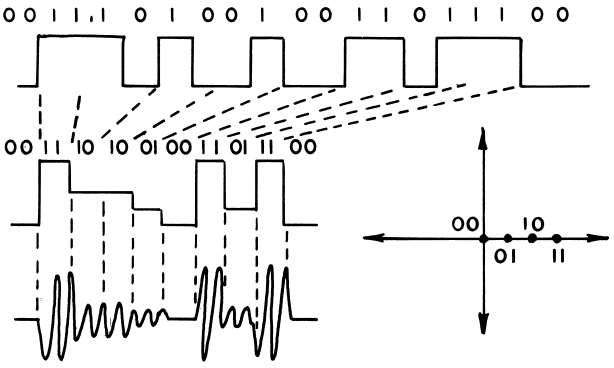


FIGURE 5 Amplitude modulation with four states and the corresponding signal-space diagram showing how the four amplitudes are coded in terms of magnitude and phase.

received signal is rectified and smoothed using a low-pass filter, and then the message signal is reproduced by passing the smoothed signal through a trigger circuit with the thresholds indicated.

C. Multiple Level Modulation

The preceding discussion assumed binary signals. With multiple level (M-level) modulation, it is possible to transmit more than two amplitudes, frequencies, or phases. If four levels of amplitude are used, as shown in Fig. 5, each distinguishable level can represent a dibit (two bits) of information. This doubles the effective bandwidth but only at the expense of greater susceptibility to noise. If eight levels were used, a tribit (three bits) could be sent at any one time, but the noise susceptibility would be even greater.

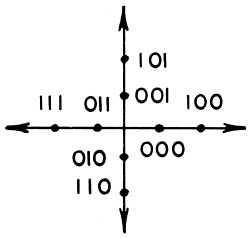
The noise susceptibility of M-level amplitude modulation can be reduced by using M-level phase modulation. Thus, dibits 00, 01, 10, and 11 could be represented by phase shifts of 90°, 0°, 180°, and 270°, respectively. M-level amplitude and phase modulation can be combined, and the result is sometimes known as *quadrature amplitude modulation* (QAM). Figure 6a shows a signal-space diagram (also known as the *signal constellation*) for a QAM system in which a combination of two levels of amplitude and four levels of phase shift (eight possible combinations) are used to transmit a tribit of information in one of eight possible (amplitude, phase) states. Figure 6b shows a more complicated modulation technique in which a quadbit (four bits) of information can be transmitted in one of 16 possible combinations of phase and amplitude modulation. Note that this scheme is much more complex than the others described previously. A synchronizing signal first establishes the absolute phase. The first bit, Q_1, of each quadbit determines the signal element amplitude to be transmitted. The next three bits, Q_2, Q_3, and Q_4, determine the phase shift or change relative to the absolute phase of the preceding element.

M-level modulation techniques such as QAM are finding increasing use in modern data-transmission systems. A 9600-bit/sec QAM modem, for example, maps a quadbit of information into one of 16 (2^4) possible combinations of amplitude and phase, as shown in the space-signal diagram of Fig. 6b and transmits 2400 quadbits/sec.

A related modulation technique is trellis code modulation (TCM), in which redundant code bits are added and transmitted with the data. A 14,400-bits/sec TCM modem, for example, assembles six data bits and generates a redundant seventh bit using two of the data bits and a binary convolutional encoding scheme. The resulting seven bits are mapped into one of 128 (2^7) possible combinations of amplitude and phase, much like QAM, and 2400 of these seven-bit symbols are transmitted each second to obtain the 14,400-bits/sec effective data transmission rate. The 14,400 comes about because only six of the seven bits represent actual data.

At the receiving end, the three encoded bits (the redundant bit and the two data bits used to generate it) select one of the eight (2^3) subsets, each consisting of 16 amplitude–phase combinations, while the remaining four bits select one of the 16 combinations from the selected subset. The redundant seventh bit ensures that only certain sequences

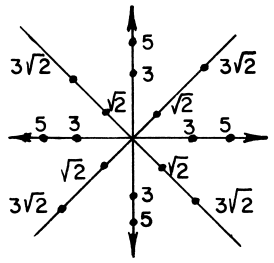
Tribit	Relative Amplitude	Phase Shift (deg)
000	1	0
001	1	90
010	1	270
011	1	180
100	2	0
101	2	90
110	2	270
111	2	180



(a)

Quadbit Q_1	Relative Amplitude	Absolute Phase (deg)
0	3	0 , 90 , 180 , 270
1	5	0 , 90 , 180 , 270
0	$\sqrt{2}$	45 , 135 , 225 , 315
1	$3\sqrt{2}$	45 , 135 , 225 , 315

Quadbits $Q_2Q_3Q_4$	Phase Shift (deg)
000	45
001	0
010	90
011	135
100	270
101	315
110	225
111	180



(b)

FIGURE 6 M-ary modulation using a combination of amplitude and phase modulation. (a) Two levels of amplitude and four levels of phase shift; (b) a more complicated technique.

of amplitude–phase combinations are valid and that any two valid sequences of combinations are far apart on the signal-space diagram to help in the decoding.

The major differences between QAM and TCM are the encoder that generates the redundant bit and the resulting signal-space diagram which, for TCM, has twice as many amplitude–phase combinations because of the introduction of the redundant bit. By careful choice of the signal-space diagram and the convolution code to generate the redundant bit, TCM can have a noise tolerance of twice that of QAM (see Payton, 1985) which translates to higher transmission rates over bandwidth-limited channels. Typical modems using QAM are limited to 9600 bits/sec while newer TCM modems can operate reliably at speeds up to 33,600 bits/sec over telephone channels. Newer so-called 56,000 bits/sec modems are asymmetric in that they can typically receive data at up to 56,000 bits/sec but transmit at only up to 33,600 bits/sec using different modulation techniques for the transmit and receive directions. Because such modems can use data compression, they can achieve effective throughput rates two or more times the above rates on compressible files.

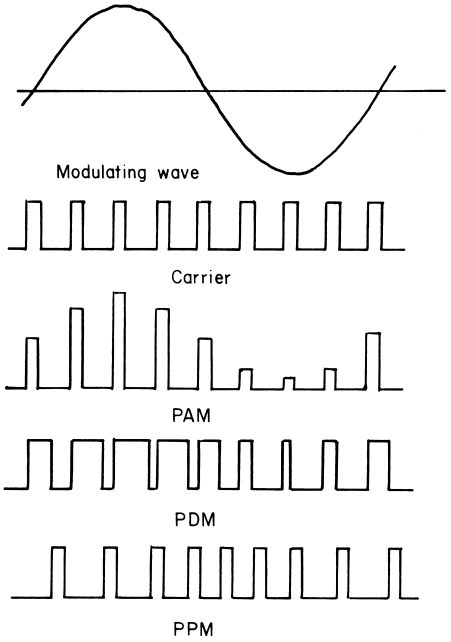


FIGURE 7 PAM, PDM, and PPM using a sinusoidal modulating wave and a pulse carrier.

D. Pulse Modulation

All the modulation techniques described previously use a continuous sinusoidal carrier wave that is modulated by analog or digital signals. Because of the continuous carrier wave, these are sometimes referred to as continuous-wave (CW) modulation techniques. In pulse modulation, the carrier is not a continuous wave but a periodic pulse train whose amplitude, duration, or position is varied in accordance with the message. Pulse amplitude (PAM), pulse duration (PDM), and pulse position (PPM) modulation are illustrated in Fig. 7. Note that PPM consists of equal-width pulses derived from the trailing edge of PDM pulses. PPM has an advantage over PDM since the latter can require significant transmitter power for transmitting pulses of long duration.

E. Pulse Code Moudulation

In pulse amplitude modulation, the amplitude of the pulse can assume any value between zero and some maximum value. Pulse code modulation (PCM) is derived from PAM but is distinguished from the latter by two additional signal-processing steps, called *quantizing* and *encoding*, that take place before the signal is transmitted. Quantizing replaces the exact amplitude of the samples with the nearest value from a limited set of specific amplitudes. The sample amplitude is then encoded, and the codes are transmitted typically as binary codes. This means that, unlike other modulation techniques described so far, in PCM both the sampling time and the amplitude are in discrete form.

Representing an exact sample amplitude by one of 2^n predetermined and discrete amplitudes introduces an error called *quantization noise* that can be made negligible by using a sufficiently large number of quantizing levels. Studies have shown that using 8 bits per sample to represent one of 256 quantizing levels provides a satisfactory signal-to-noise ratio for speech signals (see Rey, 1983). The sampling rate is usually determined from the sampling theorem, which states that a baseband (information) signal of finite energy with no frequency components higher than W Hz is completely specified by the amplitudes of its samples taken at a rate of $2W$/sec. The corollary of the sampling theorem states that a baseband analog channel can be used to transmit a train of independent pulses at a maximum rate that is twice the channel bandwidth W. These results are important in determining appropriate sample rates and bandwidths for conversion between analog and digital signals.

Applying the sampling theorem to speech signals that are limited to 4000 Hz, we find that they need to be sampled 8000 times/sec to be completely specified. Using PCM with 8 bits to represent one of 256 discrete amplitude samples, 8×8000 or 64,000 bits/sec are required to transmit the 4000-Hz voice signal. If we now use the corollary to the sampling theorem, we find that a channel with a bandwidth of 32,000 Hz is required to transmit the 64,000 bits/sec needed to specify the 4000-Hz voice

signal. Although it is true that PCM requires more bandwidth than the baseband analog signal (32,000 Hz bandwidth for the 4000-Hz voice signal in the above example), this is more than offset by the following:

1. PCM has very high immunity to noise.
2. PCM repeater design is relatively simple.
3. The PCM signal can be completely reconstructed at each repeater location by a process called *regeneration*.
4. PCM provides a uniform modulation technique suitable for other signals on many different types of media including wire, coaxial cable, free space, and optical fibers.
5. PCM is compatible with time division multiplexing.

F. Comparison of Modulation Techniques

The various modulation techniques that have been described have distinct advantages and disadvantages in terms of cost, immunity to noise, and other impairments commonly encountered in communication channels. Amplitude modulation (AM) is widely used for radio programs. It is easy to maintain and relatively low in cost, but it is susceptible to noise. Frequency modulation is more effective in terms of noise tolerance and more suited for data transmission than AM. Phase modulation is more complex and costly but is relatively immune to noise and theoretically makes the best use of bandwidth for a given transmission rate. Various forms of phase and hybrids of phase and amplitude modulation (QAM and TCM) are increasingly used for data communication over analog channels at rates up to 56,000 bits/sec. Although it requires higher bandwidth, PCM has the advantage that the signal is regenerative and has greatest immunity to noise. Optical fibers, with their very high bandwidth, are well suited for PCM. Consequently, fibers and PCM are rapidly becoming the two leading technologies for transmission of data and digitized analog signals.

VII. MULTIPLEXING

Multiplexing is a technique that allows a number of lower bandwidth communication channels to be combined and transmitted simultaneously over one higher bandwidth channel. At the receiving end, demultiplexing recovers the original lower bandwidth channels. The main purpose of multiplexing is to make efficient use of the full bandwidth of a communication channel and achieve a lower per channel cost. The three basic multiplexing methods in use are space-division multiplexing, frequency-division multiplexing, and time-division multiplexing.

A. Space-Division Multiplexing

Space-division multiplexing refers to the physical grouping together of many individual channels or transmission paths to save physical space. A large number of wire pairs, coaxial cables, and/or optical fibers are usually grouped together to form a larger cable, such as the one illustrated in Fig. 11a. Each wire pair, fiber, or coaxial cable in the main cable is a communication channel giving a high total aggregate bandwidth. With each such individual channel in the cable capable of being frequency- or time-division multiplexed, such cables have enough bandwidth to carry more than 100,000 two-way voice channels in a cable diameter of under 3 inches.

B. Frequency-Division Multiplexing

As illustrated in Fig. 8, frequency-division multiplexing (FDM) divides the frequency spectrum of a higher bandwidth channel into many individual smaller bandwidth communication channels. Signals on these channels are transmitted at the same time but at different carrier frequencies. Guard bands are needed between the frequency channels to reduce interchannel interference.

Perhaps the most familiar example of FDM is radio broadcasting. Stations broadcast continuously but are assigned and occupy a different frequency in the radio spectrum. Program signals use amplitude or frequency modulation to modulate a carrier whose frequency is allocated to that station. Tuning circuits in the radio receiver select a given frequency and allow the signal from one station to be separated from the others. The signals are in the form of modulated electromagnetic waves using the atmosphere as the communication channel.

Frequency-division multiplexing for voice signals over telephone channels is very similar. At one end, a number of modulators with carrier frequencies differing by 4000 Hz modulate the various channels onto the higher bandwidth channel. At the receiving end, an equal number of demodulators tuned to the same frequency bands as the modulators receive and demodulate the multiplexed signal into the corresponding channels. This is

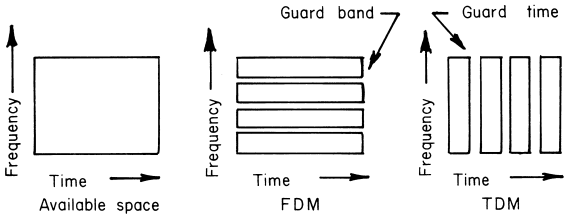


FIGURE 8 Relationship between frequency and time in FDM and TDM.

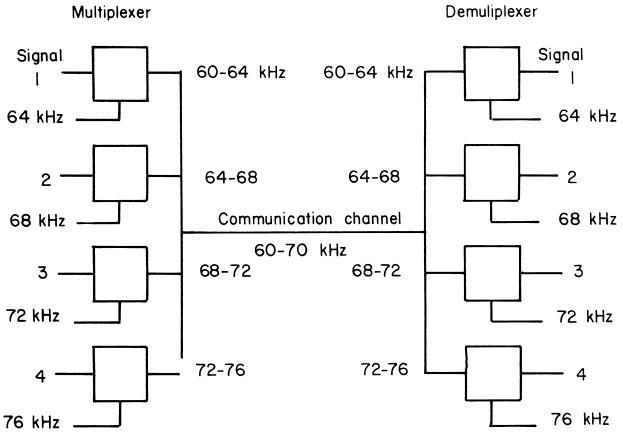


FIGURE 9 Frequency-division multiplexing

illustrated in Fig. 9 for one-way transmission, for the sake of simplicity.

C. Time-Division Multiplexing

Time-division multiplexing (TDM) operates by giving the entire channel bandwidth to a stream of characters or bits from each of the low-speed channels for a small segment of time, as illustrated in Fig. 8. This allows many low-bandwidth channels to be accommodated on a high-bandwidth channel by interleaving them in the time domain. Guard times are used to separate time slices, and the transmitting and receiving ends must be synchronized. A familiar example of a particular type of TDM might be the input–output bus of a computer servicing many peripherals for short periods of time, one at a time.

D. Multiplex Hierarchies

As networks grow and get more complex, hierarchies of multiplexing are required whereby low-bandwidth channels are multiplexed onto high-bandwidth channels, which in turn are multiplexed onto even higher bandwidth channels, and so on. In the FDM hierarchy, multiplex levels correspond to increasingly higher frequency bands. In the TDM hierarchy, multiplex levels correspond to increasingly higher pulse rates. For example, using PCM, each voice channel requires 64,000 bps. Twenty-four voice channels can be time-division multiplexed onto a T1 carrier operating at 1.544 Mbps, while four T1 carriers can be multiplexed onto a T2 carrier operating at 6.312 Mbps. A T4M carrier operates at 274.176 Mbps and can carry traffic from 42 T2 carriers or up to 4032 voice channels.

When hierarchies of multiplexing are used, it is important to have accurate carrier frequencies (pilot tones) for FDM and stable timing signals for TDM to synchronize transmitting and receiving stations.

VIII. TYPES OF DATA TRANSMISSION MEDIA

Data transmission media provide the physical communication channel to interconnect the transmitting and receiving ends. Media that constrain and guide or conduct communication signals are called *bounded* or *guided media*. Media that permit signals to be transmitted but do not guide them are called *unbounded media*. Bounded media include open wires or other conductors, twisted wire pairs, coaxial cables, waveguides, and optical fibers which are conductors of light. The atmosphere and outer space are examples of unbounded or unguided media for transmitting broadcast radio, television, terrestrial microwave radio, and satellite communications. Table II summarizes the various media, the main applications, the typical frequency bands or bit rates used, and the relative frequency of use.

It should be noted that the multiplexing hierarchies described in Section VII.D imply hierarchies in the transmission media. A single intercontinental voice or data message, for example, may use the following hierarchy of transmission media:

1. Twisted pair between source and the local telephone exchange

TABLE II Commonly Used Transmission Media

Medium used	Main application	Typical frequency range	Relative use
Open conductor	Very short distances	Varies with distance	Very high
Twisted wire pair	Up to 50 miles	Varies with distance	Very high
Coaxial cable	Short to long haul	3–500 MHz	High
Waveguide	Short distances	3–12 GHz	Low
Optical fiber	Short or long haul	Up to 16 Gbits/sec	High
Atmosphere/microwave radio	Short haul	10.7–11.7 GHz	Medium
	Long haul	3.7–6.425 GHz	High
Space/satellite	Long haul	5.925–6.425 GHz up; 3.7–4.2 GHz down	Medium
		14.0–14.5 GHz up; 11.7–12.2 GHz down	High
Atmosphere/broadcast radio	Short to long distances	Varies with distance	Very high

2. Microwave transmission between the local exchange and the central office which switches lines onto higher capacity trunks and vice versa
3. Optical fiber cable between the central office and the toll office which switches the continental trunk to an overseas trunk
4. Satellite transmission or optical fiber between toll offices on the two continents
5. Optical fiber from the intercontinental toll office to the overseas local exchange
6. Twisted pair between the overseas local exchange and the final destination

IX. TRANSMISSION IMPAIRMENTS

In practice, communication signals are degraded by physical limitations of the transmission media such as bandwidth, impairments arising from within the channel such as echo, and impairments introduced from outside the channel such as impulse noise. The impact of such impairments depends on the type of signal transmitted and whether the channel is designed for analog or digital transmission. They can attenuate, amplify, or severely distort the signal as it passes through the communication channel and introduce errors. Impulse noise, which can be caused by electrical storms, for example, has little effect on the reception of analog speech signals because of amplitude limiters in the channel and the relative tolerance of the human ear to occasional errors. However, such noise can obliterate digital data signals.

The planning, design, installation, operation, and maintenance of communication channels for transmission of analog or digital signals depend upon an understanding of the impairments to which these channels are susceptible and the effect of these impairments on the signal. Some commonly encountered impairments are briefly described below. Other impairments related to specific transmission media will be described in later sections.

A. White or Gaussian Noise

White noise has a normal (Gaussian) amplitude distribution and causes the background hiss occasionally encountered over telephone or radio channels.

B. Impulse Noise

Impulse noise usually consists of a short-duration (less than 100 to 200 milliseconds), high-amplitude burst of noise energy that is much greater than normal peaks of message circuit noise. The short duration does not unduly impair analog speech signals but can obliterate digital data signals. For this reason, data signals are usually transmitted in blocks containing error-detecting codes so that when errors are detected, the data can be retransmitted. Because of its relative short duration, impulse noise results in greater error rates at higher data rates since at higher rates it becomes increasingly difficult to distinguish between a data and noise pulse. Impulse noise is usually caused by lightning storms and voltage transients by electromechanical switching systems. The objective for telephone lines is to have no more than 15 impulse counts in 15 minutes for 50% of all telephone communications.

C. Cross-Talk Noise

Cross-talk is usually caused by capacitive or inductive coupling between adjacent channels. It can be a problem in high-speed digital circuits. If it is intelligible on analog channels (i.e., a conversation on an adjacent circuit can be overheard), it becomes particularly objectionable, not only because it impairs the conversation but also because it creates a loss of privacy. Methods to control cross-talk include shielding conductors, separating tightly coupled circuits, impedance matching of lines, and suppressing nonlinearities. Telephone companies have a goal of limiting the cross-talk index, which is the actual percentage probability of receiving an audible, intelligible speech signal on a call, to under 0.5% on most of their channels.

D. Quantizing Noise

Quantizing noise is caused by errors introduced when quantizing an analog signal into one of 2^n discrete amplitudes. It can be controlled by using a code with a suitably large value of n (e.g., 8 for voice signals). A codec is a coder-decoder that is used in PCM systems for analog-to-digital and digital-to-analog conversion.

E. Impedance Mismatch

If long wires, wire pairs, and coaxial cables are not terminated in their characteristic impedance, reflections can take place, which can cause high standing-wave ratios that appear as noise in the system. This can be overcome by terminating long lines with a load impedance that is equal to or very close to the characteristic impedance of the line, as explained in Section XII.

F. Attenuation Distortion

Attenuation distortion (sometimes called *amplitude/frequency distortion*) occurs when the relative magnitude of different frequency components of a signal are altered during transmission over a channel. A transmission

medium that is ideal has a flat frequency response for the range of frequencies over which the spectrum of the information signal is nonzero. The distortion is caused by capacitive and inductive reactances in series and in parallel with the electrical conductors comprising the channel. Attenuation distortion is much less a problem in voice transmission than in data transmission because of the redundant nature of speech signals as opposed to the small amount of redundancy in data signals, where the loss or alteration of one bit usually alters the meaning of the code word in which the bit is contained.

G. Envelope-Delay Distortion

A signal that carries information has a phase component in addition to amplitude and frequency. To obtain an undistorted signal, it is required that the transmission medium have not only a flat amplitude-versus-frequency characteristic but also a linear phase-versus-frequency characteristic. As in the case of attenuation distortion, typical transmission media do not exhibit ideal phase response because of distributed stray inductance, capacitance, resistance, and conductance which affect the time or phase relationship between various frequency components of a transmitted signal. This is known as *envelope delay*, or *group distortion*, and causes various frequency components of a complex signal to propagate through the transmission medium at different velocities, causing the composite signal at the receiver to be distorted. If the distortion is sufficiently large, late-arriving energy from one pulse can interfere with the start of the next pulse, resulting in a phenomenon known as *intersymbol interference*. Equalizers are devices that attempt to equalize the envelope delay and compensate for the different components that make up amplitude distortion.

H. Frequency Translation

Frequency translation is a phenomenon that results in a frequency shift whereby all frequency components in the modulated signal are shifted by a constant amount. It is generally due to oscillator drift or frequency offset in the carrier wave equipment. Frequency translation can be especially troublesome in systems that transmit FDM signals. This type of system relies heavily on channel filters in the modulation–demodulation process. Since frequency translation causes a shift in the energy spectrum for which the system was designed, the spectrum shift causes some of the desired information-signal energy to encounter the undesirable amplitude and phase-distortion characteristics that are usually found at the band edges of filters. If the frequency error is sufficiently large, the system is seriously degraded, especially for high-rate data transmission. This is because many modems use carriers that are synchronized between the receiver and transmitter, and frequency errors in the media appear as errors in the modems. On telephone channels, the objective is to hold the frequency shift to ± 2 Hz for acceptable performance.

I. Phase Jitter

Phase jitter refers to small changes in the phase of the received signal and appears as a slight frequency modulation on the carrier or baseband signal. It often takes the form of multiples or submultiples of ac power frequencies. It is usually caused by reactive coupling between equipment associated with power lines or jitters in signals used for timing (clocks). Phase jitter is of little consequence in voice transmission since the human ear is not sensitive to small changes in phase or frequency. However, phase jitter can seriously deteriorate data transfer since it causes the data pulses to jitter, and, if large variations occur, then one data pulse may try to occupy the time slot of another and cause an error, especially in TDM systems. In addition, phase jitter may make synchronization of the transmitter and receiver more difficult.

J. Harmonic Distortion

Harmonic distortion is usually caused by clipping or limiting the transmitted signal, which causes higher harmonics of the transmitted signal at the receiver. In most cases, this is not a serious problem.

K. Echo

Talker echo is a telephone line impairment that is caused by reflections when long lines are not terminated in their characteristic impedance. It causes part of the speaker signal to be reflected back and can seriously interfere with the talker's speaking process. If the elapsed time (i.e., the delay) is long and the echo path loss is inadequate (i.e., there is not enough loss to attenuate the echo sufficiently), the interference may be so great that speaking is nearly impossible. Impedance mismatches are usually caused by hybrids in the signal path. Hybrids are circuits used for converting a two-wire circuit to a four-wire circuit and vice versa. Echoes can be controlled using echo suppressors, which detect echoes and automatically insert a loss over the listener's speaking path to prevent the echo from reaching the talker. When both speak at the same time, the suppressor inserts a loss in both directions, resulting in undesirable clipping of speech signals. An echo canceler, rather than inserting a loss in the return path, uses the transmitted speech signal to generate a replica of the echo and subtracts this signal from the echo, thus canceling it.

Using echo suppressors and cancellers reduces the echo to tolerable levels on 99% of all telephone connections in which this problem is encountered (Rey, 1983).

L. Inductive Interference

Because power and telephone companies tend to serve the same customers and share the same right of way, telephone wires are exposed to electromagnetic fields created by power distribution systems. If the telephone lines are not perfectly balanced, a voltage difference between the two conductors of a cable pair may develop in the presence of a strong electromagnetic field that can result in an audible noise level. This type of inductive interference is sometimes known as *electromagnetic interference*, or EMI for short. Telephone sets and transmission media are designed to have high tolerance of 60-Hz hum, but higher harmonics can cause problems. This is minimized by balancing lines, grounding cable shields, and separating power and telephone media where necessary. As a last resort, magnetic core devices are used to minimize this type of noise.

M. Message Circuit Noise

Message circuit noise is a signal composed of noise from a number of sources such as power hum, inductive interference, Gaussian noise, central office noise, and impulse noise, which have been previously discussed. It causes impairments similar to those caused by its various component parts.

N. Ground Noise

Ground noise is caused by voltage differences between the receiving and transmitting ends due to current flow through the ground return path. Transmission media consisting of a single wire over a ground, as shown in Fig. 10a, are especially susceptible to this type of noise.

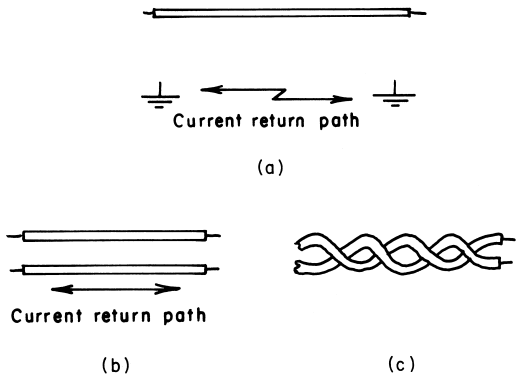


FIGURE 10 Single and paired cables. (a) Single cable over ground; (b) open pair; (c) twisted pair.

X. BOUNDED TRANSMISSION MEDIA

Bounded transmission media constrain and guide communication signals. They include single-wire (or single conductor) lines, paired cable, coaxial cable, optical fibers, and waveguides. The first three media are electrical conductors, optical fibers are conductors of light, and waveguides constrain and guide electromagnetic waves.

A. Single-Wire Lines

Single-wire lines use a single wire or some other electrical conductor to provide a path for an electric current with an electrical ground providing the return path, as shown in Fig. 10a. Since this is the medium used almost exclusively to interconnect components in all types of electronic systems over short distances (most connections between components on a circuit board or even on a silicon chip are examples of a single-wire-over-ground transmission medium), a substantial portion of Section XII will be devoted to this medium, with emphasis on its proper use in circuits where high bandwidths are involved.

B. Paired Cable

Because of ground noise problems associated with a single conductor when transmitting signals over longer distances, two-conductor paired cable is used, with the second cable providing the return path for the signal current. The two cables can be parallel to each other (open wire pair) or twisted (twisted pair), as shown in Figs. 10b and c. Open wires are susceptible to cross-talk and electromagnetic interference from power lines. As explained in Section XII, twisted pairs reduce the distance between the conductors and reduce the inductive cross-talk by reducing the area of the current loop. Consequently, twisted pairs have replaced open pairs for telephone transmission except in some rural areas. To further reduce susceptibility to noise and electromagnetic interference, cable pairs can be shielded. In many cases, space-division multiplexing is used by combining many twisted pairs into a larger cable. Twisted pairs are stranded into a ropelike form called a *binder group*, and several binder groups are twisted together around a common axis to form the cable core. A protective sheath is wrapped around the core, resulting in a cable like that in Fig. 11a. Paired cable is made in a number of standard sizes and may contain from two to several thousand cable pairs.

Paired cable is the main type of medium for local telephone and data transmission. It can be installed easily using commonly available tools. Today, it is typically used for relatively short distances (normally less than 10 miles). It has bandwidth limitations and is susceptible to noise.

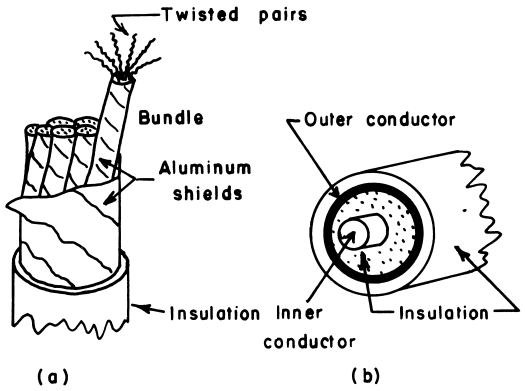


FIGURE 11 (a) Multipair cable and (b) coaxial cable.

As frequency or bit rates of transmission increase, the repeaters needed to amplify or regenerate the signals must be placed closer together to overcome the increased effect of the transmission impairments. With proper design, twisted pair cables can carry about 12 voice channels per pair with repeaters every 2 to 4 miles.

Table III briefly summarizes the various categories of unshielded twisted-pair cabling systems commonly used today and their main applications. Most new building cabling today consists of category 5 cable because it can be used for both telephone service and local area networking at up to 100 Mbits/sec. New standards are currently being discussed for higher category cables to extend local area network support up to 1000-Mbits/sec or gigabit ethernet.

C. Coaxial Cable

Coaxial and twinaxial (two coaxial cables in a single jacket) cables are used for analog or digital transmission requiring high bandwidth. As shown in Fig. 11b, a coaxial cable consists of an inner conductor completely surrounded by an outer conductor, with high-quality insulation between the two. The outer conductor is surrounded by a protective sheath. As with twisted pairs, coaxial cables can be bundled together with twisted pairs to form a larger cable with very high total bandwidth.

Coaxial cables have various characteristic impedances, with 50, 75, and 93 ohms being the most common. They have the advantage of operating at very high frequencies, which allows them to carry a very large number of analog or digital channels using FDM. Since the outer conductor is grounded and provides an effective shield that improves with increasing frequencies, the cross-talk between adjacent coaxial cables decreases with frequency, rather than increasing as is the case with unshielded cable pairs.

TABLE III Categories of Twisted Pair Cabling Systems

Category	Maximum data rate/(Mbits/sec)	Typical application
1	<1	Old telephone service
2	4	IBM token-ring networks
3	16	10 Mbits/sec ethernet
4	20	IBM 16 Mbits/sec token ring
5	155	100 Mbits/sec ethernet, 155 Mbits/sec ATM

In the telephone network, coaxial cable was used primarily on intercity routes for the long-haul network where heavy traffic existed. The Bell L5 coaxial cable, for example, had a capacity of 10,800 voice circuits per carrier channel with repeaters every mile. In these applications, the installed cost per voice circuit is lower than for paired cable. Old intercontinential submarine cables also used coaxial cables, but in this environment the design, operational, and reliability requirements are different from cables used on land. More recently, coaxial cables were being used for local area networks within or between buildings. Such networks interconnect a large number of workstations and computers using baseband transmission with TDM or broadband transmission with FDM. They have a low incidence of errors, support high data rates and can be tapped into easily, and devices such as taps, controllers, splitters, couplers, multiplexers, and repeaters are readily available.

Coaxial cables are also used extensively in community antenna television (CATV) or cable TV systems for distribution of television and music programs. However, coaxial cables for voice and data are being rapidly replaced by other media. Specifically, modern local area networks tend to use category 5 (or better) twisted pairs for intrabuilding cabling and optical fibers for interbuilding cabling. Similarly, coaxial cables carrying voice signals have or are being replaced by optical fibers, since the latter are much more cost effective because of their high bandwidth capability.

D. Waveguides

A waveguide is a rectangular or circular pipe, usually made of copper, that confines and guides very high-frequency electromagnetic waves between two locations. Compared with coaxial cable, it has a very low attenuation at microwave frequencies, which is its main advantage. Its main disadvantages are that it must be manufactured with extreme uniformity to achieve the low attenuation and great care must be taken during installations to minimize sharp bends, which also increase attenuation. These disadvantages have limited the application of waveguides primarily to carrying 3- to 15-GHz-range signals from the base of microwave radio towers to the dish antennas at the top.

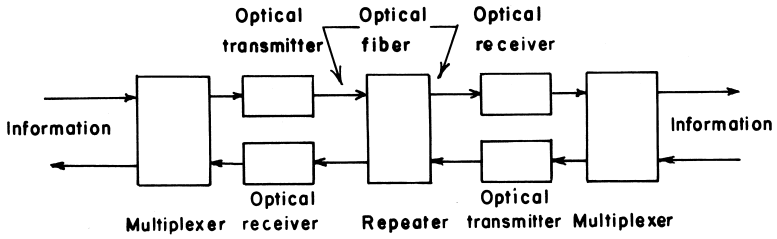


FIGURE 12 Typical optical fiber communication system.

E. Optical Fibers

Optical fibers have become the preferred medium for terrestrial communication because they can carry gigabits of information per second over short or long distances. The basis of fiberoptic, or lightguide, transmission systems is an optical fiber, which is a thin, flexible glass or plastic fiber through which light is transmitted. An optical fiber is actually a waveguide that guides the propagation of optical frequency waves directly or through total internal reflection at the fiber boundaries. A fiberoptic transmission system is shown in Fig. 12 and is similar to a conventional transmission system, except that the transmitter uses a light emitting diode (LED) or a laser diode (LD) to change electrical signals into light signals while the receiver uses a photodiode or similar device to convert the light signals back into electrical signals.

Compared with wire, twisted pairs, or coaxial cable, optical fibers offer advantages so great in terms of information-carrying capacity and signal protection that they have replaced many of the older media used in the last decade. These advantages include the following:

1. *Large bandwidth.* Optical fibers with suitable optical transceivers have bandwidths of tens of GHz compared with an upper limit of 500 MHz for coaxial cable. These bandwidths are achievable because the information-carrying capacity of a transmission medium increases with the carrier frequency and for optical fibers the carrier is light, which has frequencies several orders of magnitude higher than the highest radio frequencies. With currently available systems, a single fiber cable can accommodate 150,000 voice-frequency circuits, and this will increase rapidly with the introduction of new optical fiber technologies.
2. *Low losses.* As shown in Fig. 13, the attenuation of an optical fiber is essentially independent of the modulation frequency, whereas copper wires and coaxial cables exhibit increasing attenuation with increased frequency. The advantage of the flat attenuation response of low-loss optical fibers is that repeaters need not be placed so close together. While a 44.7-Mbit/sec T3 carrier in a digital telephone system has a standard repeater spacing of 1.1 km for coaxial cable and 6.4 km for older optical fibers, optical systems with repeater spacing of more than 100 km are being planned using new, low-loss fibers. This means that interoffice trunks will require no repeaters since the distances involved are usually

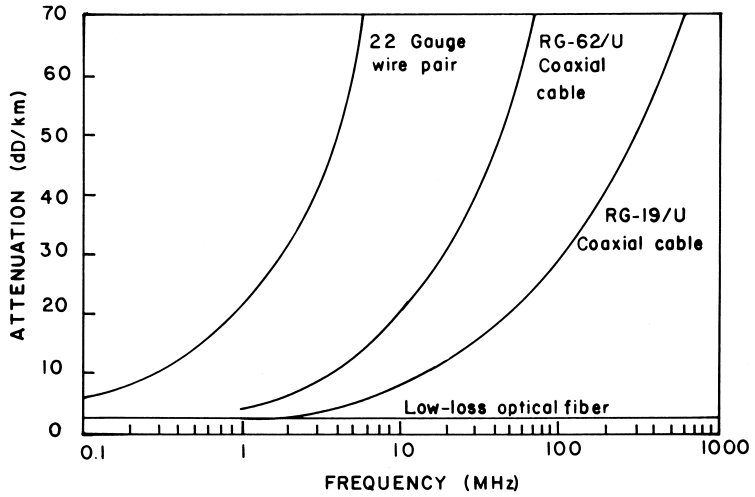


FIGURE 13 Attenuation as a function of frequency for some commonly used media.

shorter than the above repeater spacing for new fibers.

3. *Electromagnetic interference* (*EMI*) *immunity*. Because optical fibers are made of insulating materials such as glass and plastic, they are not affected by ordinary electromagnetic fields. With conductors such as copper wire, shielding is required in many cases to reduce the effects of stray magnetic fields. The EMI immunity of optical fibers also offers the potential for signal protection through low error rates of 1 per 10^{10} bits or better. This has the potential for reducing the cost and overhead associated with the complex error checking and error-correction mechanisms that have to be used with other transmission media.
4. *Small size and light weight*. A fiber with a 0.062-mm diameter and a 3.5-mm protective jacket has far more information-carrying capacity and can easily replace a cable consisting of 900 twisted copper wire pairs with an 8-cm outer diameter. The smaller size for a given capacity and the fact that glass and plastic weigh less than copper mean that fibers are lighter by a factor of 10 to 100 than copper cables of equivalent capacity. Both size and weight are important factors when considering overcrowded conduits running under city streets.
5. *Security*. Unlike electrical conductors, optical fibers do not radiate energy. Consequently, eavesdropping techniques that can be used on the former are useless with fibers. Furthermore, it is very difficult to tap an optical fiber since, in general, this affects the light transmission enough to be detected.
6. *Safety and electrical isolation*. Because fiber is an insulator, it provides electrical isolation between data source and destination. This avoids ground noise often encountered when systems are interconnected by electrical conductors. Furthermore, fibers present no electrical spark hazards and can be used in applications where electrical codes and common sense prohibit the use of electrical conductors.

Despite their many important advantages, optical fibers have several disadvantages. They are a relatively new technology and not yet understood well by enough designers, who are comfortable designing with electrical conductors and who base their solutions on what they know best. This situation is improving rapidly as industry-wide standards related to this new technology continue to emerge; the costs of light sources, detectors, fibers, and other related hardware continue to fall with increased use; and more technicians are trained to handle, install, and maintain optical fiber systems. Since it has been established that this technology can outperform cable as a transmission medium in many situations, the next decade should see a great increase in its use.

1. Types of Fibers

Optical fibers are made of plastic, glass, or silica. Plastic fibers are the least efficient but tend to be larger, more economical, and more rugged. Glass or silica fibers are much smaller, and their lower attenuation loss makes them more suited to high-capacity channels. A third type of fiber, with size and performance intermediate to the other two types, has a glass or silica core with a plastic cladding.

The basic optical fiber consists of two concentric layers—the inner core and the outer cladding, which has a refractive index lower than the core. The construction is such that light injected into the core always strikes the core-to-cladding interface at an angle greater than the critical angle and is therefore reflected back into the core. Since the angles of incidence and reflection are equal, the light wave continues to zigzag down the length of the core by total internal reflection, as shown in Fig. 14a. In effect, the light is trapped in the core, and the cladding not only provides protection to the core but also may be thought of as the "insulation" that prevents the light from escaping. The exact characteristics of light propagation depend on fiber size, construction, and composition, as well as on the nature of the light source. Although fiber performance can be reasonably approximated by considering light as rays, more exact analysis must deal with field theory and solutions to Maxwell's equations, which show that light does not travel randomly through a fiber but is channeled into modes that represent allowed solutions to the field equations. Consequently, it is more informative to classify fibers by refractive index profiles and the number of transmission modes supported.

The two main types of refractive index profiles are step and graded. In a step-index fiber, the core has a uniform refractive index n_1 with a distinct change or step to a lower

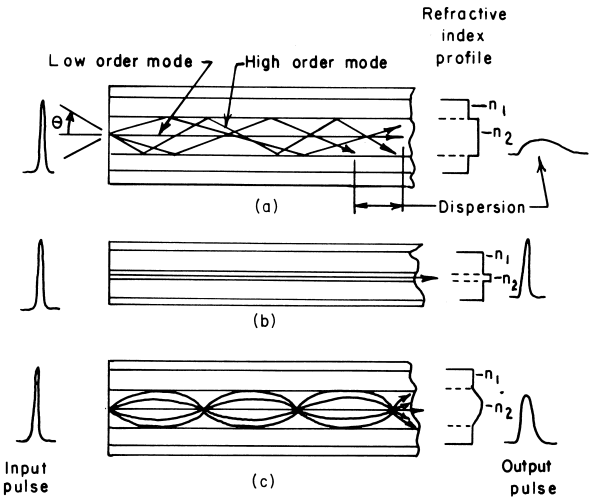


FIGURE 14 Characteristics of common optical fibers. (a) Multimode step index; (b) single-mode step index; (c) multimode graded index.

index n_2 for the cladding. In a graded index fiber, the refractive index of the core is not uniform but is highest at the center and decreases until it matches that of the cladding. The index profiles and the light-wave propagation supported by these different fibers are illustrated in Fig. 14.

The simplest type of fiber is the multimode step-index fiber, which usually has a core diameter from 0.05 to 1.0 mm. The relatively large core allows many modes of light propagation, and since some rays follow longer paths than others, their original relationship is not preserved. The result is that a narrow pulse of light has a tendency to spread as it travels down the fiber, as shown in Fig. 14a. This spreading of light pulses is known as *modal dispersion* and is 15 to 40 nsec/km for typical multimode step fibers. Consequently, such fibers tend to be used for transmission over short to medium distances.

Modal dispersion can be reduced or eliminated by using a fiber with a core diameter small enough that the fiber propagates efficiently only the lowest order mode along its axis, as shown in Fig. 14b. Single-mode, step-index fibers typically have core diameters between 0.002 and 0.01 mm and are very efficient for very high-speed, long-distance transmission. Their small size, however, makes them relatively difficult to work with, especially when fibers must be linked.

Modal dispersion can also be reduced by using graded-index fibers, as shown in Fig. 14c. The core is a series of concentric rings, each with a lower refractive index moving from the center of the core to the cladding boundary. Since light travels faster in a medium of lower refractive index, light farther from the fiber axis travels faster. This means that in a graded-index fiber, high-order modes have a faster average velocity than lower order modes; therefore, all modes tend to arrive at any point at nearly the same time. Light is refracted successively by the different layers of the core, with the result that the path of travel appears to be nearly sinusoidal. In such fibers, modal dispersion is typically well under 10 nsec/km and sometimes even less than 1 nsec.

As in the case of electrical conductors, optical fibers are protected by enclosing many fibers in a protective sheath, which is usually made of some tough plastic material. Because the fibers are so thin, the cable is usually strengthened by adding steel wire or Kevlar aramid yarn with a high Young's modulus to give the cable assembly flexibility and tensile strength. Cables are also available containing both optical fibers and electrical conductors, with the latter usually being used to provide power for remotely located repeaters.

2. Signal Degradation

Signal degradation in optical fiber systems is caused by one or more of the following:

1. *Attenuation*. Attenuation, or transmission loss (dimming of light intensity), is caused by absorption and scattering. Absorption is the optical equivalent to electrical conductor resistance and is usually caused by fiber impurities that absorb light energy and turn it into heat. The amount of absorption by these impurities depends on their concentration and the light wavelength used. High-absorption regions, such as that occurring in the 950-nm wavelength range due to hydroxyl ion (OH^-) impurities, should be avoided, for example. Scattering results from imperfections in fibers. Rayleigh scattering comes from the atomic and molecular structure of the core and from density and composition variations that are introduced during manufacturing. Unintentional variations in core diameter, microscopic bends in the fiber, and small discontinuities in the core-to-cladding interface also cause scattering loss. For commercially available fibers, attenuation ranges from about 1 db/km (decibel per kilometer) for premium small-core fibers to over 2000 db/km for large-core plastic fibers. Since attenuation losses vary with the light wavelength used, it is important for optical performance to match carefully the characteristics of the light-source transmitter and the fiber.

2. *Dispersion*. Dispersion, a measure of the widening of a light pulse as it travels along the fiber, is usually expressed in nsec/km. Dispersion limits the bandwidth or information-carrying capacity of a fiber, since input pulses must be separated enough in time that dispersion does not cause adjacent pulses to overlap at the destination to ensure that the receiver can distinguish them. Dispersion can be considered as the optical analog of envelope delay distortion in electrical conductors. Dispersion can be of two types. Modal dispersion arises from the different length of paths traveled by the different modes. Material dispersion is due to different velocities of different wavelengths. In single-mode fibers, which exhibit no modal dispersion, material dispersion is the sole frequency-limiting mechanism. High-quality optical fibers have a dispersion of about 1 nsec/km or even less.

3. *Other losses*. Interconnection is a critical part of optical fiber communication links. Fibers must be connected or spliced to provide low-loss coupling through the junction. Precise alignment results in low loss, but the small size of fiber cores together with dimensional variations in core diameter and alignment mechanisms make this a formidable task. Factors affecting these losses include:

- Differences in core diameters of fibers to be joined
- Differences in the numerical aperture of fibers to be joined; the numerical aperture (NA) is a measure of the light-gathering capability of a fiber and is equal to $\sin\theta$ where θ is the half-angle of the cone within which all incident light is totally internally reflected by the fiber core, as shown in Fig. 14a.

- Core-to-cladding eccentricity, which can cause misalignment
- Core ellipticity, which can reduce the contact area
- Lateral misalignment of the axes of the two cores to be joined
- Angular misalignment of the two cores
- Fiber-end separation
- Tilting angle between the two cores if the fiber is not cut on an axis that is perfectly perpendicular to the core axis

A splice, which refers to a permanent interconnection, has a typical loss of about 0.1 db. A connection refers to detachable or temporary interconnections, which in practice have less precise alignments, and results in a loss of up to 4 db using inexpensive connectors for large fibers.

3. Light Sources and Detectors

In an optical fiber communication system, the light source must efficiently convert electrical energy (current and voltage) into optical energy in the form of light. A good source must be

1. Small and bright, to permit the maximum transfer of light into the core of the fiber
2. Fast, to respond to rapidly changing modulating signals encountered in high-bandwidth data systems
3. Monochromatic (i.e., produces light within a narrow band of wavelengths), to limit dispersion within the fiber
4. Reliable, with a lifetime in the tens of thousands or preferably in the hundreds of thousands of hours of operation

The most commonly used light sources are gallium arsenide light emitting diodes (LEDs) and laser diodes (LDs). These devices are small, with sizes compatible with the cores of fibers, and emit light wavelengths in the range of 800–900, 1300, and 1530–1560 nm. In general, LEDs are not as monochromatic as LDs; consequently, they tend to be used for applications that require bit-rate capacities of several hundred Mbits/sec or less. They have a typical lifetimes of 100,000 hours and a light-output power characteristic that is almost linear over a large range of driving current and hence can be used for transmitting analog signals using amplitude modulation. LDs produce light with a much narrower spectrum than LEDs and can transfer more power into the fiber. This combination of characteristics allows them to be used in systems working at Gbit/sec rates, and even at hundreds of Gbits/sec using cavity lasers and dense wavelength-division multiplexing (DWDM).

Optical detectors convert optical energy into electrical energy. Devices most commonly used for this purpose include PIN and avalanche photodiodes (APDs). A PIN diode is a specially made diode with intrinsic material between the P and N materials to give a faster response time to incident light energy. The photodiodes are usually made of silicon because of their sensitivity in the 750- to 950-nm wavelength region. Silicon PIN diodes convert light power input to electrical current output with a quantum efficiency of over 90% and a response time of 0.5 μsec or less. APDs have much faster response times than PIN diodes and can be used at higher bit rates.

4. Modulation in Optical Fiber Systems

Although the light output power of LEDs and LDs is relatively linear over a wide range of drive currents indicating they are suitable for continuous-wave modulation (e.g., amplitude modulation), today's optical fiber communication systems are more suitable to digital rather than analog operation. In the digital mode, the light source is switched on and off, which greatly simplifies the detection process at the destination.

The selection of line-signaling format is an important consideration in optical communication systems. It is desirable to use a format that is self-clocking at the receiver and conserves output power at the source, especially when using LD sources, since they should be driven to high power output for only short intervals to improve their lifetimes. Some of the popular digital signaling formats are illustrated in Fig. 15, including:

1. Nonreturn to zero (NRZ), in which a change of state occurs only if there is a 1-to-0 or a 0-to-1 transition. A string of ones is a continuous "on" condition, while a string of zeros is a continuous "off" condition.

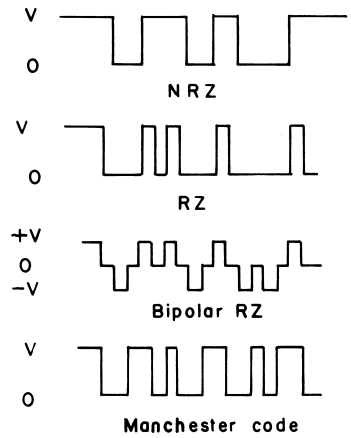


FIGURE 15 Examples of binary signal formats for optical fibers.

2. Return to zero (RZ) format, in which there is a complete pulse transmitted for each logic one condition. Note that the pulse width must be less than the bit interval to permit the return to zero condition.
3. Bipolar RZ format, in which a pulse is transmitted for each logical one and zero condition. Again, the pulse width must be less than the bit interval to permit the return to zero condition.
4. Manchester code format, in which, by convention, a logic zero is defined as a positive-going transition, while a logic one is defined as a negative-going transition occurring at times other than bit-time boundaries.

For optical systems, the RZ and Manchester code formats are good candidates because they are self-clocking and because the relatively short pulses, even for a continuous string of logical ones, conserve the life of LDs. Note, however, that these formats require at least twice the bandwidth of the NRZ format for a given bit rate. Pulse modulation techniques such as PDM and PPM are also sometimes used.

5. Wavelength-Division Multiplexing

In older conventional fiberoptic systems, a single light frequency or wavelength carries information along the fiber. Wavelength-division multiplexing (WDM), increases the information-carrying capacity of a fiber by assigning multiple incoming optical signals to specific light frequencies (or wavelengths) within a designated frequency band and multiplexing them onto a single fiber. This is the optical equivalent of conventional frequency-division multiplexing described in Section VII.B. The term *dense wavelength division multiplexing* (DWDM) is usually reserved for optical systems that use more than eight different optical wavelengths to simultaneously carry information over a single fiber.

The technologies that support WDM and DWDM include new types of laser diodes and optical amplifiers. The diodes can produce multiplicity of closely spaced wavelengths in the 1530- to 1560-nm region, with a spacing between wavelengths of 0.8 nm. The optical amplifiers consist of optical fibers doped with the rare element erbium (EDFAs, or erbium-doped fiber amplifiers) which can amplify all the wavelengths as a group when pumped by an external light source. This allows the optical signals to be boosted directly without converting them to electrical form. These technologies can support 32 channels (wavelengths) each capable of carrying 10 Gbits/sec, or a total of 320 Gbits/sec per fiber, and even faster systems are being planned.

A significant advantage of WDM is that each individual wavelength or channel can carry an independent stream of data at the same or at different bit rates as the other channels. Besides greatly improving the capacity of optical fibers, WDM therefore provides a mechanism for a unified optical communication infrastructure capable of meeting the new service demands of tomorrow.

F. Outlook for Bounded Media

Conducting wires, such as conductors over grounds, strip lines, and twisted pairs, provide a convenient mechanism for interconnecting electronic components to make up more complex systems. They are also widely used for local area networks within buildings and for connecting telephone sets to the local central office. Of all transmission media, they are the most pervasive and are likely to remain so in the near future since they are by far the least costly. Compared to other transmission media, they are also by far the most limited in terms of bandwidth, distances spanned, and susceptibility to interference. Because they radiate electromagnetic energy, they can interfere with other equipment and should not be used in an environment where data security is important. Although they will be increasingly replaced by other media for high transmission rates and distances greater than a few kilometers, they will remain the most pervasive medium for use over short distances because of their simplicity and low cost.

Coaxial cables have made inroads in high-frequency system interconnects, cable television, high-bandwidth telephone transmission systems, and local area networks. They will continue to be used for the former two applications but are rapidly being replaced by other media for the latter two applications. Specifically, most intrapremises local area networks now use category 5 (or better) twisted pairs, while optical fiber systems are rapidly replacing coaxial cable for wide area telephony and data transmission.

Major improvements in light sources, detectors, splicing technology, dense wavelength-division multiplexing equipment, and saliton transmission systems, together with the emergence of standards and reduction in component costs, make optical fiber systems the transmission medium of choice for future high-bandwidth, long-distance, terrestrial communication systems. The dominant application areas will continue to be telephone, video and data trunk lines, local area and local distribution networks where the distance and bandwidth requirements exceed the capabilities of category 5 wiring, and high-interference environments where electrical conductors would be too costly to shield. In the more distant future, optical and electro-optical technologies may be used to solve interconnection problems in high-speed digital systems, even for very short distances (Goodman et al., 1984). Today, electrical conductors are used for this purpose almost exclusively.

XI. UNBOUNDED TRANSMISSION

Bounded media require a physical connection between two points to conduct or guide current, electromagnetic waves, or light. In unbounded media, no such physical connection is required. Space or air is the transmission medium for electromagnetic waves, which are therefore unbounded by the medium. In the past, this medium was used almost exclusively for some forms of television or radio transmission, which includes broadcast, microwave, and satellite transmission. More recently, the use of this transmission medium has been increasing rapidly for voice and data communication using a variety of cell telephone and mobile computing technologies. The major advantages of using space or air as the transmission medium include:

1. No physical path is needed between the source and destination. This is extremely important since providing a voice line or cable between two points may be economically infeasible or physically impossible, as in the case of communication with a mobile computer.
2. The source (transmitter) and destination (receiver) can both be mobile.
3. Transmission can be point to point or point to many points simultaneously (multi-access) as in broadcast radio.
4. A broad spectrum from low to high bandwidth is available.
5. The method can be quickly implemented and no right of way is required for the transmission path.

Radio transmission also has some disadvantages. It is subject to interference and a number of propagation anomalies that may require considerable engineering effort to overcome. It also requires space in the crowded radio transmission spectrum and therefore may require licensing and design approval from regulatory agencies.

A. Radio Frequency Spectrum

In radio transmission, electromagnetic power from a radio transmitter is coupled by the transmitter antenna into air or free space. In radio reception, electromagnetic radio waves are intercepted by a receiving antenna and coupled into a receiver for detection. The antennas can be omnidirectional as in broadcast radio or highly directional as in microwave radio transmission. With an omnidirectional antenna, a receiver can receive signals (as well as noise and interference) from other transmitters located anywhere within receiving distance. With directional transmitting and receiving antennas, signals are transmitted in a very narrow angular sector (e.g., 1° for microwave transmission), thus a receiver can only detect signals emanating from a narrow sector. If the transmitting and receiving antennas are aimed at each other, good communication can be established with relatively low transmitted power and little interference to other radio links, a key point in conservation of the limited radio frequency spectrum shown in Table IV. Such point-to-point radio communication means that frequencies can be reused in the same general vicinity without interference. This should be contrasted with onmidirectional broadcast systems where frequencies can be reused only if transmitters using the same frequencies are located sufficiently far apart so as not to cause mutual interference at local receivers. Table IV shows some of the most common radio frequency bands in current use and their typical applications, characteristics, and principal modulation methods.

B. Signal Degradation

Radio waves, like light waves, are subject to reflection and refraction. They are also subject to attenuation losses due to atmospheric and natural phenomena such as rain, snow, and fog. These result in three major types of signal degradation: multipath interference, fading, and attenuation losses.

Reflections from the earth's surface, the ionosphere, natural or manmade objects, and atmospheric refraction can create multiple paths between the transmitting and receiving antennas. Depending on the relative path distance, the reflected wave is shifted in phase with respect to the original wave, which can cause interference at the receiver called *multipath interference*. Since the amount of phase shift is frequency dependent, the combined received signal is also frequency dependent, which can lead to serious problems in wideband transmission.

Fading is caused by abnormal changes in the refractive index of the atmosphere. Normally, the atmosphere refracts or bends radio waves back toward the surface of the earth. However, abnormal distribution of temperature, humidity, and heavy ground fog can cause radio waves to be bent toward the surface much more than normal so that they never reach the receiving antenna, causing changes in received signal strength or even complete loss. Variation or, specifically, reduction of received signal strength at different periods in time is called *fading*.

As transmission frequencies increase, path attenuation losses due to the atmosphere also increase. More serious losses are caused by fog, snow, and especially rain, which becomes very significant at frequencies above 4 GHz. The effects of these losses are fading and increased error rates. They are usually allowed for during the design process by using published meteorologicial data of the region in which a radio link is located.

TABLE IV Radio Frequency Spectrum

Frequency	Name	Characteristics	Principal modulation methods[a]		Typical uses
			Analog	Digital	
3–30 KHz	Very low frequency (VLF)	Very noisy, very limited bandwidth	Seldom used	ASK, FSK, PSK	Sonar, long-range navigation
30–300 KHz	Low frequency (LF)	As for VLF	Seldom used	ASK, FSK, PSK	Navigational aids, radio beacons
300–3000 KHz	Medium frequency (MF)	Noisy, limited bandwidth, large transmitting antenna	AM	ASK, FSK, PSK	Commercial AM radio, maritime radiotelephone, distress calls
3–30 MHz	High frequency (HF)	Noisy,long distance but subject to fading, subject to interference, moderate antenna size	AM	ASK, FSK, PSK	Shortwave radio, CB radio, ship to coast, point to point
30–300 MHz	Very high frequency (VHF)	Line-of-sight range, moderate noise and interference, small antenna size	AM, FM	FSK, PSK, others	Television, FM radar, land and air mobile traffic, police
300–3000 MHz	Ultrahigh frequency (UHF)	Line-of-sight, high bandwidth, low noise, high congestion in some areas	FM	FSK, PSK, others	UHF television, radar, space telemetry, microwave links
3–30 GHz	Superhigh frequency (SHF)	Line-of-sight, bandwidth to 500 MHz, low noise, narrow antenna beams	FM	FSK, PSK, others	Microwave links, radar, satellite communication
30–30 GHz	Extremely high frequency (EHF)	Line-of-sight, bandwidth to 1 GHz, low noise, high attenuation, very small antenna	FM	FSK, PSK, others	Radar landing systems, satellite communication

[a] ASK = amplitude shift keying; FSK = frequency shift keying; PSK = phase shift keying.

C. Terrestrial Microwave Radio

Terrestrial microwave transmission uses highly directional antennas for line-of-sight propagation paths using frequencies in the 4- to 12-GHz range. The antennas are usually parabolic with diameters ranging from 12 inches to several feet, depending upon their spacing. The spacing between repeater stations depends upon the geographical terrain over a given route, the technology used in the terminal equipment, and the transmitter power permitted by the local regulatory agency. For long-distance transmission, typical repeater spacings are 20 to 30 miles, but longer spacings are possible in areas where atmospheric conditions result in little fading.

Microwave links can carry several thousand voice channels using frequency-division multiplexing. They are an important source of competition with coaxial systems and have the advantages that no physical facility is required to guide the microwave energy between separate stations and that these stations can normally be 20 to 30 miles apart as opposed to about 1 mile for coaxial systems.

Using the earth's atmosphere as the transmission medium, however, results in multipath interference and fading. This can be minimized using frequency diversity, which means the transmission of the same radio signal over different microwave frequencies at the same time. Since, for a given set of atmospheric conditions, radio signals at different frequencies experience different degrees of multipath interference and fading, the strongest received signal is selected for retransmission or detection. Frequency diversity has the disadvantage of using more of the available radio frequency spectrum; consequently, space diversity is sometimes used as an alternative approach, especially for paths that experience severe fading. Space diversity uses two receiving antennas, usually mounted on the same tower and separated vertically by several wavelengths. By switching from the regular to the diversity antenna whenever the signal level drops, the received signal can be maintained nearly constant. Neither frequency nor space diversification is effective in countering attenuation caused by rain. The only remedy for this is the use of greater transmission power or shorter repeater spacing.

D. Satellite Transmission

Satellite transmission consists of a line-of-sight propagation path from a ground station to a communications

satellite (up link) and back to an earth station (down link). The satellite is usually placed in a geosynchronous orbit about 22,300 miles above the earth so that it appears stationary from any point from which it is visible and acts like a repeater in the sky. The ground station includes the antennas, buildings, and electronics necessary to transmit, receive, multiplex, and demultiplex signals. The frequency spectrum used is similar to that used for terrestrial microwave radio. The ground station antenna is usually highly directional, while the satellite antenna has a larger beam width to cover a larger portion of the earth's surface and to be able to communicate with many widely separated earth stations simultaneously.

The capability of a satellite to act as a repeater for many different earth stations is called *multiple access* (MA). Three main methods are currently in use to accomplish this: frequency-division multiple access (FDMA), time-division multiple access (TDMA), and code-division multiple access (CDMA). In FDMA, circuits between different earth stations are assigned different frequency bands within the allowable bandwidth. In TDMA, the entire bandwidth is allocated to each earth station for a short time, just like in time-division multiplexing. CDMA (also called *spread spectrum multiple access*) uses a pseudo-random code and operates in both time and frequency domains. It is effective against jamming techniques and is used primarily in military satellite communications.

Satellites have had a dramatic impact on communication topologies and pricing. They can be used for transmission of program video, voice, or data signals almost anywhere on earth, no matter how remote the location or whether it is fixed or mobile like a ship. They provide multichannel capabilities, wide bandwidths, and high data rates. The transmission cost is independent of the distance between the source and destination. Furthermore, only one satellite repeater is required for most transmissions. This characteristic of satellite transmission may make it superior to terrestrial systems such as microwave or coaxial cable, since the latter systems require many repeaters in tandem to cover long distances, and amplification of the signal by each repeater tends to increase the effects of distortion and noise.

Because of the greater distance between the earth and the satellite repeater, attenuation and transmission delays can cause problems. Attenuation can be overcome by using high gain, narrow beams, and path elevation angles greater than 20° for the ground antennas. The high path elevation reduces the distance the signals travel through the atmosphere to reduce attenuation and fading. However, rain attenuation can still be a problem, especially at the higher carrier frequencies, but it can be minimized by a form of space diversity since rain is unlikely to occur at two widely separated ground stations. The total transmission delay is approximately 0.5 sec. which is much higher than for terrestrial transmission media and is due to the much longer distance the signal must travel (a minimum of $2 \times 22{,}300$, or 44,600 miles). This delay can impair the quality of voice communication but has the greatest detrimental effect on data transmission unless communication protocols are designed to match the characteristics of this transmission medium. Low-orbit satellites (250 to 1000 miles above the earth) can overcome this delay problem, and such systems are beginning to be deployed even though the number of satellites required is greater and their liftimes are shorter than for geosynchronous satellites.

E. Cellular Telephones

Cellular telephones use a combination of low-powered radio transmitters and receivers (transceivers) to provide voice and data telephone service to mobile users. They include small antennas and work by transmitting information to cellular towers which cover a given area or cell. The towers within these cells are connected to a central switching station (usually by fiberoptic cable), which in turn is connected to the rest of the telephone system. Cellular calls are therefore received by the towers using the atmosphere as the transmission medium and are relayed to the telephone network. The cells overlap so that when the caller moves from one cell to another the cell towers automatically transfer the call so that the communication is uninterrupted.

Cellular phone networks cover most metropolitan areas and their coverage is rapidly expanding into rural areas. While simple in theory, reliable production systems are very complex. Older systems used the 824- to 893-MHz band, but newer systems operate at much higher frequencies (around 1900 MHz) and use PCM rather than analog modulation techniques. A good description of cellular telephone technologies can be found in Macario (1993).

F. Other Unbounded Media Transmission Systems

Communications satellites and terrestrial microwave radio are by far the most frequent systems using unbounded media for transmission of data. However, many other systems using unbounded media are available and are commonly used, especially for transmission of voice and program signals. A partial listing and a brief description of these systems is given below. A more detailed description may be found in Freeman (1981) and Pooch et al. (1983).

1. *Radio and television broadcast*. These systems use omnidirectional antennas to broadcast program signals to receivers in the local area. AM radio uses the band from

535 to 1605 KHz, while FM radio uses 88 to 108 MHz. Television uses the very high and ultrahigh frequency bands to provide the bandwidth required for acceptable transmission of video program signals. A receiver can select one of many local stations by tuning to the frequency allocated to the desired station, which is a form of FDM.

2. *High-frequency radio*. As little as 40 years ago, high-frequency (HF) radio carried almost all transatlantic telephone traffic. Today it is used primarily for short-wave radio and for telephone communications to ships at sea and to countries not connected by cable and with no satellite antenna. Long-distance communication is possible with HF radio since the frequencies employed allow the radio waves to be reflected by the ionosphere back to earth until they reach their final destination in one or more hops. For this reason it is sometimes called *ionospheric radio transmission*. Because of the movement and changes of the ionosphere, HF radio is subject to fading, interference, distortion, and periodic blackouts. Thus, it is rarely used for data transmission unless extreme precautions are taken for detection and correction of errors.

3. *Tropospheric-scatter radio transmission*. In tropospheric-scatter radio transmission, the troposphere is used to scatter and reflect a fraction of the incident radio waves back to earth. With two highly directional antennas pointed at the troposphere (which is about 6 miles up, as opposed to 30 miles for the ionosphere), enough radiowave energy is reflected back to result in very reliable over-the-horizon communication systems with effective ranges of 100 to 600 miles. Such systems are used where it is not possible or economical to use land lines or even microwave radio. Island chains or very rugged mountain ranges are good examples of terrain where such systems are used. The carrier frequencies are in the 900- to 5000-MHz range and can carry several hundred voice circuits over short distances of 100 miles, although 72 circuits is more typical over longer distances. It is subject to fading, which is usually overcome using space diversity with two receivers and transmitters at each end of the link with the antennas separated by at least 30 wavelengths to achieve reliable transmission.

4. *Packet radio*. Packet radio can be used to allow mobile computers to communicate with a server that is connected to an omnidirectional home base transceiver. Data to be transmitted to a computer are formatted into fixed-length packets, which include the address of the destination at the beginning and cyclic redundancy check characters at the end of the packet. Each computer receives all the transmitted packets and discards those that are not addressed to it. Packets with detected errors are not acknowledged by the computer, which causes the home base to retransmit the message after a certain time. When computers transmit to the home base, it is possible that two or more computers may transmit simultaneously, in which case the messages are not acknowledged. If a computer does not receive an acknowledgment within a predetermined time, it reschedules the packet for retransmission at a random time. Eventually, all packets will be received correctly. Different frequencies can be used for the server-to-computer and computer-to-server links to allow concurrent reception and transmission. A variation of packet radio can allow any workstation or computer to communicate with any other without the master–slave relationship described above.

5. *Light transmission systems*. Light-transmission systems use a light-emitting or laser diode as a transmitter and a photodiode or phototransistor as a receiver. As in microwave radio, the transmission path must be line-of-sight, and air is the transmission medium. The problem with such systems is that extreme care must be taken to align the light source and detector. Even then, problems can occur in heavy rain, fog, or snow, which attenuate the transmitted light signal and impair error-free reception. Bounded communication using optical fibers overcomes these problems.

6. *Wireless ethernet networks*. New technologies for 10- and 100-Mbits/sec ethernet are emerging to provide more effective point-to-point wireless connectivity and to support mobile computer users. A typical point-to-point wireless ethernet network may include several remote locations, each with a unidirectional antenna, connected to a central site equipped with an omnidirectional antenna that is connected to the enterprise ethernet. Similar technologies allow mobile computers with omnidirectional antennas to communicate with a local ethernet network much like cellular phones can communicate with the terrestrial telephone system. Every indication is that the use of such systems will grow very rapidly in the near future.

G. Outlook for Unbounded Transmission Media

The ease of establishing communication links will remain the primary attraction of links using unbounded transmission media. The multiple access capability of some systems using these media is also an important advantage. New developments in systems such as packet radio and wireless ethernet are of particular importance since they promise inexpensive communication between fixed or mobile workstations and computers. With satellite communications significantly reducing the cost for long-distance transmission, the main competition to systems using unbounded transmission media are optical fibers. Perhaps the

greatest obstacle to the growth of use of unbounded media is the limited frequency spectrum, but with improvements in antenna design, integrated circuits operating at high frequencies, and advances in modulation techniques to use bandwidth even more efficiently, systems using unbounded media will always play a major role in communications. An important reason for the move toward digital television is that the latter can use bandwidth more effectively and will free up much of the current television frequency spectrum to other applications.

XII. INTERCONNECTION CONSIDERATIONS FOR ELECTRICAL CONDUCTORS

It has been mentioned in Section X that electrical conductors are and will continue to be the predominant transmission medium for interconnecting components within a system. As bandwidths and bit rates increase, it becomes increasingly important to take into account the stray inductance and capacitance that influence the performance of these media. These stray reactances are strongly dependent on the geometry, conductor length, and proximity to the ground plane. By carefully positioning the conductor we can reduce one stray component, but only at the expense of increasing the other. Because of the relatively slow rate of change of stray inductance and capacitance with the distance from the ground plane, the conductor length is the major factor determining the stray reactances.

The current or voltage propagation delay T_d in a conductor is given by:

$$Td = l\sqrt{LC} \tag{8}$$

where l is the length of the conductor in meters, L is its distributed inductance, and C is its distributed capacitance per unit length. The conductor is said to be "short" if its propagation delay T_d is much less than the rise and fall times of the pulses it is expected to transmit. If its length is such that its T_d is comparable to or longer than the rise and fall times, or greater than $0.35/f$ where f is the greatest frequency per second for which it will be used, the conductor is termed as being "long," in which case transmission line theory should be used to predict its characteristics.

The transmission line parameters relevant to our discussion are the characteristic impedance and the reflection coefficient. The characteristic impedance Z_0 is the impedance seen at one end of an infinitely long line and is approximately by:

$$Z_0 = \sqrt{LC} \tag{9}$$

The reflection coefficient r is the ratio of the reflected voltage v_r to the incident voltage v_i if a conductor is not terminated in its characteristic impedance and is given by:

$$r = \frac{v_r}{v_i} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{10}$$

where Z_L is the terminating or load impedance. If a long conductor is not terminated or matched in its characteristic impedance, reflections will be present and ringing will occur if pulses with sharp leading and trailing edges are transmitted over the line. Such reflections and ringing increase the inherent electrical noise of the system in which they occur. The impedance ranges of transmission lines that are usually encountered in practice are given below. Note that a strip line is a rectangular conductor over a ground with the width of the conductor begin much greater than its thickness. This type of conductor is encountered in printed circuits, for example.

Types of transmission line	Characteristic impedance (Ω)
Wire over ground	80–400
Twisted pair	80–200
Coaxial pair	40–120
Strip line	20–140

The above relations are useful in determining the length and placement of lines. In high-impedance circuits, it is important to minimize capacitance by making the line as short as possible and placing it as far as possible from the ground. In low-impedance circuits, it is important to minimize the inductance by making the line as short as possible and placing it as close to ground as possible. For long lines, the lines should be terminated in their characteristic impedance whenever possible to minimize reflections and hence circuit noise.

The presence of stray electromagnetic fields in a system causes cross-talk voltage or current to be induced in neighboring conductors. This cross-talk may be predominantly capacitive or inductive. Capacitive cross-talk is due to stray capacitance between conductors and is predominant in circuits with large voltage swings and small currents (i.e., high-impedance circuits). The capacitive cross-talk voltage v_c is given by:

$$v_c = k_c\varepsilon(\text{voltage swing} \times \text{length of line} \times \text{function of spacing})/\text{rise time of voltage} \tag{11}$$

where ε is the dielectric constant of the medium between the conductors and k_c is a constant. In general, the closer the lines the greater the v_c. Spacing the lines farther apart and away from the ground reduces v_c but

has the effect of increasing inductive cross-talk. When the voltage swings are small and the currents are large (i.e., low-impedance circuits), inductive cross-talk usually predominates. It arises chiefly because of mutual inductance coupling between conductors and is generally given by:

$$i_i = k_i(\text{current swing} \times \text{length of line} \times \text{function of spacing})/\text{rise time of current} \quad (12)$$

where i_i is the cross-talk current, which can be reduced if the lines are farther from each other and as close as possible to the ground.

If more than one line is driven at the same time, the inductive currents in the passive line are additive. This is not the case with capacitive cross-talk. An upper bound is reached for v_c if the passive line is surrounded by active lines. Both types of cross-talk may be reduced by using special lines. Twisted pairs, coaxial lines, and multiple shielded lines all have zero mutual impedance in theory and therefore have zero inductive cross-talk. In practice, the following numbers are useful:

Type of wiring	Mutual inductance (μH/m)
"Neat" in bundles	1–7
Point-to-point over ground	0.6
Twisted pairs	0.06
Coaxial cable	0.006
Multiple shielded lines	As small as necessary

In summary, it is impossible to eliminate all the stray inductance and capacitance of conductors. It is possible only to trade one against the other by varying the characteristic impedance. In interconnecting components or systems, the goal should be to use point-to-point connections over a ground plane. All lines should be coaxial or twisted pairs and matched at least at the receiving end. Practical considerations may not permit this, but the more we approach this type of interconnections, the less trouble will be experienced with system noise and cross-talk and therefore with getting a high-speed system to operate successfully.

XIII. STANDARDS

Modern communication systems represent complex and rapidly changing technologies. Effective communications on a local, country, and worldwide basis require that communications equipment and transmission media work together in harmony. Thus, there is need for widely accepted standards. Such standards are established by a number of organizations, including the International Telegraph and Telephone Consultative Committee (CCITT), the International Organization for Standardization (ISO), the American National Standards Institute (ANSI), the Electronic Industries Association (EIA), the European Computer Manufacturers Association (ECMA), the National Communication System (Federal Standards–Telecommunications), and the National Bureau of Standards (NBS). Standards most relevant to data communication are compiled and summarized in Folts (1982).

XIV. SUMMARY

Advances in computer and communication technologies are transforming our society. They have created entire new industries and are changing the way we learn, do research, conduct business, communicate with each other, and even entertain ourselves. These changes have led to an exponential increase in communication traffic generated by the Internet, cellular telephones, faxes, modems, teleconferencing, and data and video services. Ubiquitous connectivity is removing the constraints of space and time to access services from anywhere at any time. This has led to enormous increases in the demand for additional communication bandwidth capacity, and the marketplace has responded with a variety of new technologies that make more effective use of all traditional transmission media including optical fibers, free space, and the atmosphere, as well as the older coaxial and paired cables previously used for voice communication. These technologies have created, and will continue to create, a vast global network for the rapid movement of information in the form of voice, video, and data.

SEE ALSO THE FOLLOWING ARTICLES

DIGITAL SPEECH PROCESSING • NETWORKS FOR DATA COMMUNICATION • OPTICAL FIBER COMMUNICATIONS • OPTICAL INFORMATION PROCESSING • RADIO SPECTRUM UTILIZATION • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL • VOICEBAND DATA COMMUNICATIONS

BIBLIOGRAPHY

Folts, H. C., ed. (1982). "McGraw-Hill's Compilation of Data Communications Standards," McGraw-Hill, New York.

Freeman, R. L. (1981). "Telecommunication Transmission Handbook," John Wiley & Sons, New York.

Goodman, J. W., Leonberger, F. I., Kung, S. Y., and Athale, R. A. (1984). *Proc. IEEE Special Issue on Optical Computing* **72** (7), 850–866.

Macario, M. (1993). "Cellular Radio: Principles and Design," McGraw-Hill, New York.

Payton, J., and Qurèshi, S. (1985). "Trellis encoding: what it is and how it affects data transmission," *Data Commun*. May.

Pooch, U. W., Greene, W. G., and Moss, G. G. (1983). "Telecommunications and Networking," Little, Brown & Co., Boston, MA.

Proakis, J. G. (1983). "Digital Communications," McGraw-Hill, New York.

Rey, R. F., ed. (1983). "Engineering and Operations in the Bell System," 2nd ed. AT&T Bell Laboratories, Murray Hill, NJ.

Schwartz, M. (1980). "Information Transmission, Modulation and Noise," 3rd ed. McGraw-Hill, New York.



Digital Speech Processing

Biing Hwang Juang
M. Mohan Sondhi
Bell Laboratories, Lucent Technologies

Lawrence R. Rabiner
AT&T Laboratories

I. Introduction
II. Speech Analysis and Representation
III. Speech Coding
IV. Speech Synthesis
V. Speech Recognition
VI. Speaker Verification
VII. Speech Enhancement
VIII. Concluding Remarks

GLOSSARY

Adaptive coding A coding process that adapts its attributes and process parameters according to specific characteristics of the signal that is being encoded.

Continuous speech recognition Automatic recognition of speech utterances in which words are spoken continuously without enforced pauses between words.

Formant Resonance of the vocal tract manifested acoustically as a concentration of energy in the frequency spectrum.

Isolated word recognition Automatic recognition of words spoken as an isolated sequence of words with distinct pauses between individual words.

Keyword spotting Process of automatically detecting and recognizing a key word, or a key phrase, embedded in a naturally spoken sentence.

Predictive coding A coding process that involves using past (known) values of a signal to estimate the current value of the signal. The residual (i.e., the estimation error) has much lower variance than the original signal. This fact can be utilized to improve coding efficiency.

Speaker verification Process of authenticating a claimed identity based on spoken information.

Speech analysis The process of using computational algorithms to measure properties of the speech signal.

Speech coding Process of converting a sampled speech signal into a format suitable for storage or transmission over a network. The usual motivation for speech coding is storage or transmission efficiency.

Speech enhancement Processing of speech to make it less noisy, clearer, more intelligible, or easier to listen to.

Speech production model A model for the generation of speech by the human vocal apparatus.

Speech understanding High-level processing of a sequence of spoken words that extracts meaning from the spoken input.

Subjective quality Quality of a speech signal perceived by human listeners.

Text analysis Parsing and analyzing a sequence of words to make explicit the underlying syntactic and semantic structure so as to allow proper pronunciation and grouping of the words and to facilitate automatic understanding of the meaning in the given word sequence.

Text-to-speech synthesis Process of synthesizing speech from printed, usually unconstrained text.

Verbal information verification Process of verifying a claimed identity based on spoken user-specific information, such as a password, mother's maiden name, or a personal identification number (PIN).

SPEECH is the most fundamental form of communication among humans. Digital speech processing is the science and technology of transducing, analyzing, representing, transmitting, transforming, and reconstituting speech information by digital techniques. The technology is intended to improve communication between humans as well as to enable communications between humans and machines. Communications between humans can be enhanced by digital encoding and decoding of speech for efficient transmission, storage, and privacy and by suppressing noise, correcting distortion, and compensating for hearing loss. Communication between humans and machines comprises automatic recognition and understanding of speech, to give machines an "ear" with which to listen to spoken utterances, and digital synthesis of speech, to give machines a "mouth" with which to respond back to humans.

I. INTRODUCTION

Speech is the most sophisticated signal naturally produced by humans. A speech signal carries linguistic information for the sharing of information and ideas. It allows people to express emotions and verbally share feelings. It is the most fundamental form of communication among humans. The aim of digital speech processing is to take advantage of digital computing techniques to process the speech signal for increased understanding, improved communication, and increased efficiency and productivity associated with speech activities.

The field of speech processing includes speech analysis and representation, speech coding, speech synthesis, speech recognition and understanding, speaker verification, and speech enhancement. Speech is a complex signal that is characterized by varying distributions of energy in time as well as in frequency, depending on the specific sound that is being produced. The speech signal also possesses other characteristics that make it a very efficient means for carrying semantic (meaning) as well as pragmatic (task-dependent) information. The processes of *speech analysis and representation*, which lie at the technical basis of digital speech processing, attempt to use computational algorithms to "discover," measure, and represent the important properties of speech for many applications.

One of the most important applications of digital speech processing is *speech coding*, which is concerned with efficient and reliable communication between people who may be separated by geographical distance or by time. The former forms the basis of modern telephony, which enables people to converse regardless of their locations, and the latter forms the basis for applications such as "voice mail" that let people create verbal messages at arbitrary times, and retrieve them later. Speech coding enables both telephony and voice messaging by converting the speech signal into a digital format suitable for either transmission or storage. Relevant issues in speech coding are conservation of bandwidth (the rate of the voice coder), voice quality requirements, processing and transmission delay, and processing power, as well as techniques for privacy and secure communication.

Speech synthesis is concerned with providing a machine with the ability to talk to people in as intelligible and natural a voice as possible. A speech synthesis system can be as simple as a "prerecorded" announcement machine with a limited collection of utterances, or as complicated as a full text-to-speech conversion system, which automatically converts unconstrained printed text into speech. *Speech recognition*, often viewed as the counterpart to speech synthesis, is concerned with systems that have the ability to recognize speech (literally transcribe word-for-word what was spoken), and then understand the intended meaning of the spoken words (or more properly the recognized spoken words). Speech recognition systems range in sophistication from the simplest speaker-dependent, isolated-word or phrase recognizer to fully conversational systems that attempt to deal with virtually unlimited vocabularies and complex task syntax comparable to that of a natural language system.

Speaker recognition is concerned with machine verification or identification of individual talkers, based on their speech, for authorization of access to information, networks, computing systems, services, or physical premises. At times, as an important biometric feature, a

talker's speech may also be used for forensic or criminal investigations.

Speech enhancement attempts to improve the intelligibility and quality of a speech signal that may have been corrupted with noise or distortion, causing loss of intelligibility or quality and compromising its effectiveness in communication. Speech enhancement methods are also useful in the design of hearing-aid devices, which transform the speech signal in such a way as to compensate for hearing loss in the auditory system of the wearer of the device.

Digital speech processing has continued to make substantial advances in the past decade, primarily due to the explosive growth in computational capabilities (both processing and storage), the introduction of key mathematical algorithms (see below), and the vast amount of real-world data that are being systematically made available through organized efforts among government agencies and research institutions. According to Moore's law, the computing capabilities of general-purpose processors grow at the rate of doubling the processor speed and memory every 18 months, without much increase in cost. Advances in very large scale integrated (VLSI) circuits have led to realization of extremely fast digital signal processors (DSPs) with very low power consumption. This has helped the proliferation of miniaturized devices such as cellular mobile phones. Today, a powerful DSP can run on as low as 0.9 volts, consumes a mere 0.05 mW per MIPS (million instructions per second), and has performance approaching 1000 MIPS (or GIPS). These advances in computing power are particularly beneficial in dealing with signals as complex as human speech. Finally, in the area of software, the convergence of programming languages (e.g., C and C++), the widespread use of command and interpretive languages for scientific computing such as MATLAB, and user-friendly environments for developing DSP assembly-code have made digital speech processing very accessible, practical, and easily realizable.

In the past decade, computing and digitization devices also have become ever more readily available for data recording and management. The availability of easily accessible databases has added a new direction to digital speech processing research, going beyond the traditional paradigm of hypothesis and observation made by experts in the field, to a data-driven mode in which computational algorithms are designed to learn from real-world data directly. This change of research paradigm is significant in making digital speech processing a useful practice because speech signals usually exhibit a wide range of variability, which can only be adequately observed (in a statistical sense) in a large collection of data. A human expert inherently lacks the ability to deal with such a large quantity of data. Realizing this, a number of government agencies and industrial research institutions have, in the past two decades, separately collected a number of task-specific speech databases. Such databases include speech sentences chosen for coder evaluations, the set of Naval Resource Management (RM) sentences for the study of continuous speech recognition, the set of Air Travel Information System (ATIS) sentences for speech recognition and dialog research, the set of North America News Broadcast (NAB) sentences for automatic speech transcription, and the Switchboard database for recognition of conversational telephone speech, just to name a few. These databases have guided many speech research programs and have been used for evaluating the progress in digital speech processing, thereby helping to make many system designs practically useful.

Many digital speech processing systems have been deployed for real-world services in the past decade. Major growth in the utility of voice systems exists in at least four major market sectors: telecommunications, business applications, consumer products, and government. In the *telecommunications* sector, speech coders are essential in digital and packet telephony, for traditional voice circuits as well as over the Internet. Specialized speech announcement systems use coded speech to provide timely information to customers, and speech recognition systems are used for automation of operator and attendant services as well as for retrieval of account information. In *business applications*, voice mail and store-and-forward messaging systems are in widespread use; automatic speech recognizers help people direct calls in a private branch exchange (PBX) arrangement; and voice interactive terminals let users dictate correspondence or issue voice commands for command and control of various devices or parts of a computer operating system. In the *consumer products* sector, toys incorporating speech synthesis and recognition have been available for years, and talking appliances and alarm announcement systems are also beginning to appear in household appliances. In the area of *government communications*, anticipated uses of speech processing include coding for secure communications, speech recognition for command and control of military systems, and automatic speech understanding and summarization for intelligence applications. These examples, by no means exhaustive, illustrate the burgeoning applications of digital speech processing and point to a growing market in the coming years.

II. SPEECH ANALYSIS AND REPRESENTATION

The traditional framework for analyzing speech is the source-tract model first proposed by Homer Dudley at

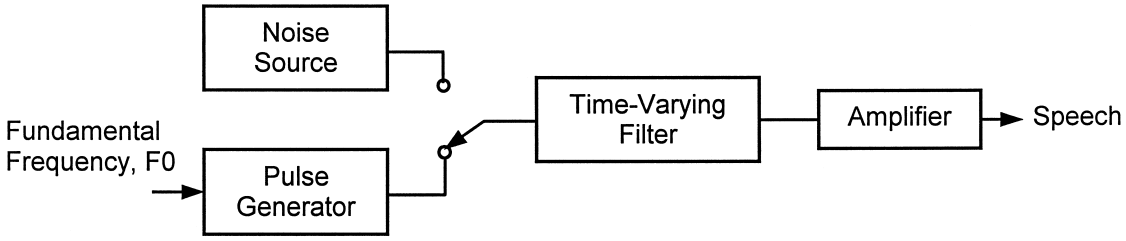


FIGURE 1 The speech production model, the basis for speech analysis.

Bell Laboratories in the 1930s. In this model, as depicted in Fig. 1, a speech excitation signal is produced by an excitation source and processed by a filter system that "modulates" the spectral characteristics of the excitation signal based on the shape of the vocal tract for the specific sound being generated. The excitation source has two components—a "buzz" source and a "hiss" source. The signal produced by the buzz source is a sequence of pulses with a controllable repetition rate (the fundamental frequency, or F0) and provides the carrier for voiced sounds such as /a/ in f*a*ther, /e/ in m*e*t, and /o/ in hell*o*. The hiss source produces a noise-like signal and provides the carrier for unvoiced sounds such as /s/ in *s*ell, /sh/ in *sh*out, and /k/ in *k*itten. The source signal is then filtered by a time-varying linear system with adjustable frequency response. The signal at the output of the time-varying filter is the final speech signal that has a time- and frequency-dependent distribution of energy. This time-varying filter system, which models the effects of the human vocal tract, is realized by a bank (ten channels in Dudley's original embodiment) of bandpass filters that span the range of speech frequencies. Any desired vocal tract frequency-response characteristic may be realized by adjusting the amplitudes of the outputs of the bandpass filters. Dudley was able to demonstrate (very successfully at the 1939 New York World's Fair) that a machine based on the production model of Fig. 1 can produce sounds that are very close in quality to that of natural speech.

Based on the work of Dudley, the source-tract model of speech production has been the dominant framework in speech analysis and representation for the past 60 years. Key elements in the model—namely, specification of the frequency response of the slowly time-varying vocal tract filters and characteristics of the source signal—are thus the targets of speech analysis, which aims at finding optimal parameter values that best define the source-tract model to match a given speech signal.

To obtain the (time-varying) frequency-response characteristics of the vocal tract filters, the class of well-known methods of spectral estimation is normally employed. One such method uses a bank of bandpass filters, each followed by a nonlinearity, to measure the power level at the output of each channel. This method can be used to generate a local estimate of the spectral profile of the speech sound at a given time. Digital filter-bank design techniques are very well understood and are relatively straightforward to implement. Alternatively, one can use digital spectral analysis methods, such as the discrete Fourier transform (DFT) or fast Fourier transform (FFT) methods, to implement the filter banks according to a prescribed structure and spectral matching criterion. The resulting spectral profile of the time-varying speech signal, when plotted in the time–frequency plane using the density of the print to indicate the corresponding power level, is called a *sound spectrograph, sonogram,* or *spectrogram.* Figure 2 shows an example of a spectrogram corresponding to a speech segment of 2.34 seconds' duration.

Another important method for analyzing and representing the time-varying vocal tract frequency response is known as *linear prediction*. Linear prediction, or linear predictive coding (LPC), is motivated by the fact that the speech signal, at any time instant, can be approximated

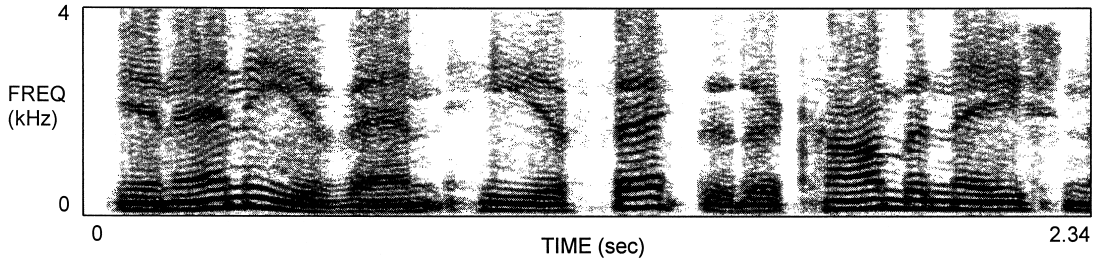


FIGURE 2 An example of a spectrogram for a speech segment.

by a linear combination of its past values. The difference between the predicted value and the true sample value is called the *residual,* or *prediction, error.* There are three important properties of LPC that make this method of analysis the most important one in most speech processing systems. First, the resulting set of optimal weights (the so-called linear prediction coefficients) that achieve the best prediction, in the sense of minimizing the residual for each short-time window of speech, defines an all-pole filter that best characterizes the behavior of the vocal tract. Second, the residual signal, when used as the excitation signal to drive the optimal tract filter, produces a signal essentially identical to the original signal. The residual signal has properties closely resembling those of the excitation source: It resembles a noise-like signal for unvoiced sounds and a pulse train for voiced sounds. Third, the optimal residual signal has a much smaller dynamic range than that of the original speech signal and is thus a preferred signal for coding (see Section III).

To model the time-varying characteristics of the speech signal, the LPC analysis procedure is repeated every few hundredths of a second. This process is generally referred to as *short-time spectral analysis.*

Another important property of the speech signal is the manner in which the vocal tract is excited. The excitation can be either voiced (in which case the vocal cords are vibrating and produce a periodic excitation) or unvoiced (in which case the vocal cords are not vibrating and a noise-like excitation is generated by air rushing through a narrow constriction somewhere in the tract). For voiced sounds, the key characteristic of the source is the fundamental frequency, which is the frequency of vocal-cord vibration. Another important speech parameter is the status of speech/non-speech activity (i.e., an indication of the presence or absence of speech).

There are several ways to determine whether a segment of speech is voiced and to determine the periodicity if it is voiced. In the time domain, one effective way is to compute the autocorrelation function of the speech signal. If the autocorrelation function exhibits no strong peak, then the speech segment is unvoiced. If a strong peak is observed, then the lag at which it occurs indicates the fundamental period of the signal (which is the reciprocal of the fundamental frequency). Another method for estimating the fundamental frequency of voiced sounds, which capitalizes on the regularity in the spectrum, is the so-called method of cepstral analysis. The cepstrum is the (inverse) Fourier transform of the log spectrum of the speech signal. Regularity in the log spectrum, such as repetitive spikes at essentially equal spacing (corresponding to the fundamental frequency), results in a clear spike in the cepstrum at a location (the "quefrency" index) corresponding to the pitch period. The cepstrum is also often used as a feature measurement in speech recognition algorithms.

III. SPEECH CODING

The goal of speech coding is to transform a sampled speech waveform into a digital representation that allows efficient transmission and storage of the signal. Such a transformation is performed by an algorithm or device called a *coder.* After storage or transmission, an approximation of the original signal may be resynthesized by means of a decoder. The process of coding and decoding, in general, introduces a processing delay and also results in a certain loss of fidelity. Generally, to achieve higher coding efficiency (i.e., lower bit rate) one must accept lower speech quality and longer processing delay. The three dimensions—coding efficiency, processing delay, and quality—are interrelated. The constraint of acceptable cost or algorithmic complexity also influences the achievable efficiency. The trade-off between these aspects must be adjusted according to the demands of a specific application. For instance, a delay of even a few seconds might be quite acceptable if the coded speech is to be stored and then retrieved at a later time. On the other hand, for telephony between humans, a total delay (due to processing plus that due to transmission) of over 200 msec would make fluent two-way communication quite difficult.

Quality of speech is a subjective measure, often expressed in terms of the so-called mean opinion score (MOS). The MOS is obtained by averaging the quality judgment from a pool of human listeners in response to a set of stimuli (speech samples) produced by the coder/decoder being tested. The subjective quality judgment is usually measured on a five-point scale ranging from 1 (unacceptable) to 5 (excellent). An MOS of 4 or above is generally regarded as high quality. A descriptive taxonomy of speech quality is as follows:

- Broadcast quality (FM bandwidth of ~7 kHz with MOS greater than 4)
- Toll quality (telephone bandwidth of ~3.2 kHz with MOS around 4)
- Communications quality (military and mobile radio with MOS around 3)

Speech intelligibility is another measure that may be used to evaluate coding/decoding algorithms, particularly those algorithms that attempt very low bit rates. Intelligibility is often tested by a diagnostic rhyme test (DRT) or diagnostic alliteration test (DALT). These tests measure errors made by human listeners between similar monosyllabic words. The DRT uses rhyming word pairs (e.g., *m*et vs. *n*et) and the DALT uses alliterative word pairs (e.g., fla*t* vs fla*k*). The words are chosen so as to uniformly cover six major linguistic attributes (e.g., voicing, nasality, etc.). Speech processing may introduce distortions that lead to increased confusions between such similar words

as judged by human listeners. Except in some rare applications, speech processing algorithms need to be able to maintain a minimum DRT score of 90% to be acceptable.

Analysis methods in speech coding are generally based on two approaches. A *waveform coder* encodes the speech such that when the signal is reconstituted at the receiver for playback, the original waveform is reproduced as faithfully as possible. A *model-based speech coder*, or *vocoder*, transforms the speech into a representation that models human speech production. These two approaches represent the two ends of the technological spectrum that covers most speech coder designs.

The most widely used method of waveform coding is called pulse coded modulation (PCM), which was historically motivated by transmission applications (and hence the term modulation). In linear PCM, the amplitude of each sampled speech signal is approximated by a binary integer (i.e., an integer expressed in terms of 0s and 1s) that takes on one of a number of values uniformly spread across the amplitude range. To be able to faithfully reproduce a speech signal without detrimental perceptual effects, at least 4096 values are needed, which means a 12-bit linear PCM representation. To take advantage of the nonuniform amplitude distribution of the speech samples, a linear PCM system is generally preceded by amplitude compression during encoding and followed by amplitude expansion during decoding. Two such nonlinearities are commonly used in today's digital telephony—namely, the A-law and μ-law companders (compressor-expander). The resulting so-called A-law or μ-law PCM is more efficient than a straightforward linear PCM, and requires around 8 bits per sample to achieve a similar quality. For telephony applications, the speech signal is usually sampled at 8 kHz and encoded with 8-bit companded PCM, with a resulting transmission rate of 64 kb/sec.

Speech is a slowly varying signal and displays substantial correlation among the values of adjacent samples. A coding method that makes use of this correlation is thus generally more efficient than simple PCM encoding. The simplest way to utilize the adjacent-sample correlation in speech is the scheme of differential PCM (DPCM), which does not encode the speech signal directly but rather the difference between the current and the previous samples. When the coding scheme is made adaptive to certain time-varying properties of the signal (rather than remaining fixed for all time, independent of the signal characteristics), still higher efficiency is possible. The method of adaptive DPCM (or ADPCM) allows toll-quality transmission of speech at 32 kb/sec. Note that these simple differential schemes incur only a single sample (0.125 msec) of delay.

More sophisticated coding schemes, such as the adaptive predictive coding (APC) method, use more previous speech data samples to better predict the current speech sample value, resulting in a residual signal of much reduced variance that is easier to encode. The idea of using a linear weighted combination of the previous sample values to predict the current value gained tremendous attention in the late 1960s and early 1970s (see Section II). This is the method of linear predictive coding (LPC), in which a set of optimal weight coefficients (the so-called predictor coefficients) are calculated at a regular rate, say once every 20 msec, to minimize the variance (or energy) of the residual within a short window of speech samples. Studies in articulatory modeling were able to relate the predictor to the vocal tract shape and its resonance structure (or the formants), enabling LPC to become one of the most successful methods in digital speech processing.

The concept of model-based coding has its root in the original source-tract framework of Homer Dudley. In the source-tract model, key components are the parameters that define the tract filter, and the pitch and voicing information that defines the excitation signal. In the encoder, these components are obtained by applying speech analysis techniques to the signal and the resulting parameters are quantized with efficient coding schemes for transmission. The speech signal may be resynthesized from these parameters (the quantized spectrum and the quantized excitation signal), which vary relatively slowly with time. By transmitting these data at a slower rate (normally less than 30 data values once every 20 msec) than the original sampled waveform (normally 8 sample values every msec), the resulting vocoder thus provides a potential for speech transmission with a much higher efficiency. The advent of LPC makes analysis and representation of the time-varying spectral shape of the corresponding vocal tract straightforward and reliable. LPC-based vocoders operating at 2.4 kb/sec have been in military deployment for secure communications since the 1970s. At this bit rate, the coarse buzz–hiss model is usually used to represent the excitation, and the quality of the resynthesized speech is quite limited with an MOS normally below 3 and a DRT score in the low 90% range.

Many new digital speech coders fall into the category of a hybrid coder, in which the concept of vocal tract representation is integrated with a waveform-tracking scheme for the excitation signal to achieve a quality suitable for telephony applications. These include the multipulse excitation LPC coder, the code-excited LPC (CELP) coder, and the low-delay CELP (LD-CELP) coder. These coders normally operate at bit rates ranging from 4.8 to 16 kb/sec.

Another major advance in speech coding, namely vector quantization, took place in the late 1970s. The basic idea of vector quantization is to encode several samples (i.e., a vector), as opposed to a single sample (i.e., a scalar), at a time. Shannon's communication theorems provided

strong motivations for the use of vector quantization. Furthermore, a number of algorithms were developed to automatically design the collection of reconstruction vectors (the codebook) according to the data distribution, so as to minimize the distortion incurred as a result of coding. Many recent speech coders employ vector quantization. When applied to model-based coding, vector quantization vocoders allow transmission of digital speech at 800 b/sec and below, with acceptable speech quality for special communications needs (e.g., in certain military applications).

For speech coding to be useful in telecommunication applications, coding systems have to be interoperable, thus the coding scheme has to be standardized before deployment. Speech coding standards are established by various standards organizations such as the International Telecommunications Union (ITU), the Telecommunications Industry Association (TIA), the Research and Development Center for Radio Systems (RCR) in Japan, the International Maritime Satellite Corporation (Inmarsat), the European Telecommunications Standards Institute (ETSI), and some government agencies.

Figure 3 illustrates the performance of various speech coders in terms of the perceptual quality of the decoded speech at the corresponding operating bit rate. The companded PCM coder at 64 kb/sec, the ADPCM coder at 32 kb/sec, the LD-CELP coder at 16 kb/sec and the algebraic-coded CELP (A-CELP) coder at 8 kb/sec are all capable of achieving an MOS of more than 4, thus are readily usable in telecommunication applications. The CELP at 4.8 kb/sec has an MOS of slightly less than 4 and is useful in several applications such as secure voice communication. The LPC vocoder at 2.4 kb/sec can only achieve an MOS of around 2.5 and is so far used only in military applications. A new multiband excitation LPC coder (ME-LPC), in which the characteristics of the excitation are determined for each frequency band separately, is able to operate at 2.4 kb/sec, and yet provides a perceptual quality approaching that of CELP at twice the bit rate. The chart also shows that the current challenge, in terms of the rate-distortion performance of a coding scheme, lies at and below 4 kb/sec, where the perceptual quality drops radically.

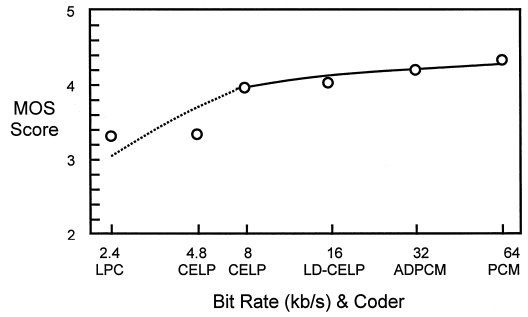


FIGURE 3 Speech quality in terms of MOS for various coders at typical bit rates.

The advent of the Internet and the possibility of packet telephony using the Internet Protocol (IP), the so-called voice-over IP (or VoIP), gives rise to another challenge in speech coding. The added dimensions from the new packet networks (as opposed to the traditional synchronous telephony network hierarchy) are

1. Network delay and delay jitter, which vary substantially depending on the network traffic conditions.
2. Packet loss due to congestion created by multiple services running on a common infrastructure.

These new applications call for a coding scheme that performs robustly even under these potentially adverse conditions.

IV. SPEECH SYNTHESIS

Modern speech synthesis is the product of a rich history of attempts to generate speech by mechanical means. The earliest known device to mimic human speech was constructed by Wolfgang von Kempelen over 200 years ago. His machine consisted of elements that mimicked various organs used by humans to produce speech—a bellows for the lungs, a tube for the vocal tract, a side branch for the nostrils, etc. Interest in such mechanical analogs of the human vocal apparatus continued well into the twentieth century. In the latter half of the nineteenth century, Helmholz and others began synthesizing vowels and other sonorants by superposition of harmonic waveforms with appropriate amplitudes. A significantly different direction was taken by Dudley in the 1930s with his discovery of the carrier nature of speech, and its corollary, the source-tract model shown in Fig. 1. By using a keyboard to control the time-varying filter and the choice of excitation for the system shown in that figure, he was able to synthesize quite good-quality, fluent speech.

The first *digital* speech synthesizer was demonstrated at Bell Laboratories around 1967. In some sense, that synthesizer was a throwback to von Kempelen's machine, with one major difference. Instead of *manipulating* a mechanical model, the synthesizer *computed* what a mechanical model would do if it were implemented. On the basis of rules derived from a study of human speech production, a computer program computed the sequence of shapes that a vocal tract would have to go through in order to generate speech corresponding to any text presented as an input. From these shapes, and the knowledge of the appropriate acoustic excitation (periodic pulses for voiced sounds,

noise-like excitation for unvoiced sounds), the program solved the wave equation with appropriate boundary conditions to compute the acoustic pressure at the lips. Finally, an electrical signal with the same waveform as the computed pressure was applied to a loudspeaker to produce the desired speech.

This method of speech synthesis has a strong appeal because it mimics the way a human produces speech. However, in spite of considerable effort, it has not yet proven possible to automatically generate good-quality speech from arbitrary text by this method. This is because it has not been possible to derive rules that work correctly in all circumstances. So far, good synthesis by rule seems to require too frequent *ad hoc* modification of the rules to be useful in practice. Also, it is difficult to generate certain speech sounds (e.g., bursts in sounds like /k/ of *k*itten) by rule.

Modern text-to-speech synthesis (TTS) is based on a much less fundamental but much more effective procedure called *concatenative synthesis.* Basically, a desired speech signal is assembled with "units" selected from an inventory compiled during a training phase of the synthesizer. The units are acoustical representations of small sub-word elements (e.g., phonemes, diphones, frequently occurring consonant clusters, etc.). The representation itself can take several alternative forms (e.g., the LPC along with the residual as described earlier). The complete process of concatenative synthesis, from text input to speech output, consists of several steps. These are outlined in Fig. 4 for one such system for the synthesis of English. The details would differ for other languages, but the general framework would be similar for a large class of languages.

Starting with the input text in some suitable format (e.g., ASCII) the text is first normalized. Normalization consists of detecting blank spaces, sentences, paragraphs, and capital letters and converting commonly occurring abbreviations to normal spelling. The abbreviations include symbols (such as $, %, &), as well as titles (e.g., Mr., Mrs., Dr.), abbreviations for months (Jan., Feb., etc.), and so on. Note that, as illustrated in the figure, the same abbreviation (Dr.) can give rise to different letter sequences (Doctor in the first example and Drive in the second), depending on the context.

The next step shown in Fig. 4 deals with the problem that a given string of letters must be pronounced differently depending on the part of speech or the meaning. So, for instance, "lives" as a verb is pronounced differently from "lives" as a noun, and "axes" is pronounced one way if it is the plural of "axe" and another way if it is the plural of "axis." There are many such confusions that the parser needs to disambiguate.

At the output of the syntactic/semantic parser, the text has been normalized and the intended pronunciation of all the words has been established. At this point, the pronunciation dictionary is consulted to determine the sequence of phonemes for each word. If a word does not exist in the dictionary (e.g., a foreign word or an unknown abbreviation), then the synthesizer falls back on letter-to-sound rules that make a best guess at the intended pronunciation.

Once the phoneme sequence has been established, the synthesizer specifies prosodic information—that is,

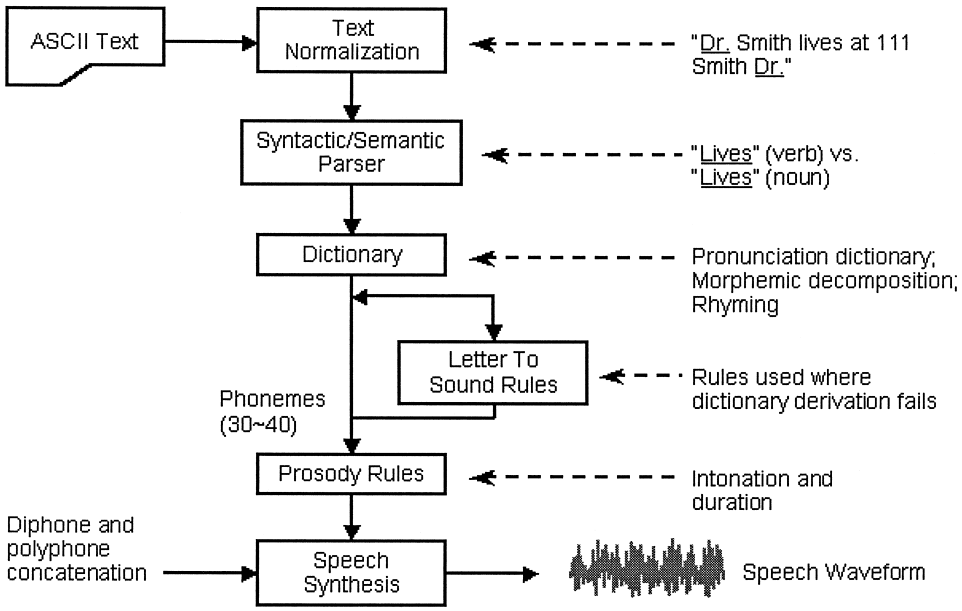


FIGURE 4 Block diagram of an English text-to-speech system.

information regarding the relative durations of various speech sounds and the intensity and pitch variations during the course of the utterance. Such information is critical to the generation of natural-sounding speech. If the wrong words are stressed in a sentence, or if the durations are inappropriate, the speech sounds unnatural. Not only that, inappropriate stress can alter the meaning of a sentence. Similarly, without a natural variation of pitch, the utterance might sound monotonous. Also, inappropriate pitch variation might, for instance, change a simple declaration into a question.

The final box in Fig. 4 does the actual synthesis of the speech. Its task is to select the appropriate sequence of units from the inventory; modify the pitch, amplitude, and/or duration of each unit; and concatenate these modified units to produce the desired speech waveform.

Currently, there are several directions in which concatenative text-to-speech synthesis is being extended and improved. One major effort is concerned with the collection of the inventory of units. Until recently, because of considerations of computational complexity and memory requirements, inventories had one (or at most a few) token for each needed unit. This token was then modified in pitch, amplitude, and duration as required by the context. Considerable effort was spent in the optimal selection of the tokens. More recently, the trend has been to have very large inventories in which most units might appear in many contexts, and with many different pitches and durations. With such an inventory, units can be concatenated after much less modification than was needed with the earlier inventories. Of course, much larger memory is required, and also the problem of searching the large inventory requires much more computation. Fortunately, both available memory and processor speed are increasing at a very rapid rate and becoming quite affordable.

Another direction is towards "multilingual" TTS. Clearly, the synthesis of a given language will have many features not shared with other languages. For example, word boundaries are not marked in Chinese; the grouping and order in which the digits of a given number are spoken in German are very different from those in English; the number and choice of the units inventory need to be different for different languages; and so on. However, rather than creating a collection of language-specific synthesizers, multilingual TTS aims at developing more versatile algorithms. The ultimate aim is that all language-specific information should be in tables of data, and all algorithms should be shared by all languages. Although this ideal is unlikely to be achieved in the foreseeable future, its pursuit focuses research more on language-independent aspects of synthesis.

Another direction of interest is that of adapting synthesis to mimic the speech of a specified person. Clearly, speech synthesized in the manner described above will sound somewhat similar to that of the person who contributed the units in the inventory. Can one make it sound like the speech of any specified person? This appears to be highly unlikely. However, it might be possible to make the synthesizer sound like several different voices by manipulating the prosodic rules and making systematic modifications of the given inventory.

We conclude this section by noting that concatenative synthesis is by no means to be considered as the ultimate in speech synthesis. Although the quality of speech generated by this method is better than that of other methods, it is by no means an accurate mimic of human speech. Ultimately, we believe, accurate modeling of the human vocal apparatus might still be necessary to achieve a significantly higher level of speech quality.

V. SPEECH RECOGNITION

Speech recognition by machine in a limited and strict sense can be considered as a problem of converting a speech waveform into words. It requires analysis of the speech signal, conversion of the signal into elementary units of speech such as phonemes or words, and interpretation of the converted sequence in order to reconstruct the sentence or for other linguistic processing such as parsing and speech understanding. Applications of speech recognition include a voice typewriter, voice control of communication services and terminal devices, information services such as voice access to news and messages, and price inquiry and order entry in telecommerce, just to name a few. Sometimes the area of speech recognition is extended to include "speech understanding," because the utility of a speech recognizer often involves understanding of the spoken words in order to initiate a certain service action (for example, routing a call to an operator).

Until the 1970s, speech recognition was mostly considered to be a speech analysis problem. The fundamental belief was that if a proper analysis method were available that could reliably produce the identity of a speech sound, speech recognition would be readily attainable. Researchers in acoustic phonetics in the past advocated this deterministic view of the speech recognition problem, citing such examples as, "A stitch in dime saves nine" (in contrast to "A stitch in time saves nine"), which they believed can be recognized correctly only via the use of acoustic-phonetic features. This view may be appropriate in a microscopic sense (e.g., to distinguish the *isolated* distinction between /d/ and /t/ in the above example) but does not address the macroscopic question of how a recognizer should be designed such that on average (in dealing with all input sounds), it achieves the lowest error rate.

The introduction of the statistical pattern-matching approach to speech recognition, which matured in the 1980s and continues to flourish into the 21st century, helps set the problem on a solid analytical basis. Statistical pattern matching and recognition are motivated by Bayes' decision theory, which asserts that a pattern recognizer requires a knowledge of the statistical variation in the observations in order to be able to achieve the lowest recognition error probability. Error probability means, essentially, how likely it is that the recognizer will make a mistake, on average, over all unknown observations (i.e., samples within the recognition test set). The knowledge of statistical variation is expressed in the form of a model, whose parameters have to be learned from real data. These data, called the *training set*, are usually a large collection of utterances with known (or manually labeled) identities. A model is a probability measure and can be considered a typical "pattern" with associated statistical variances. The process of learning the statistical regularities as well as the associated variation from the data is often referred to as "training" the recognizer. Since the error probability, or error rate, is the most intuitive and reasonable measure of the performance of a recognition system, modern speech recognizers use Bayes' theory as the basic design principle.

The most successful model used for characterizing the statistical properties of speech today is the hidden Markov model (HMM). A hidden Markov model is a doubly stochastic process. Locally in time, a sequence of observations of a speech signal is described by one of a set of random processes. Globally, the statistics of the transitions between these random processes is described by a finite-state Markov chain. This model has been found to be quite suitable for the speech signal, which indeed displays two levels of variation, one due to the uncertainty in realization of a particular sound and the other due to the uncertainty in sequential changes from one sound to another. Today, most if not all of the automatic speech recognition systems in the field are based on the hidden Markov model technique.

Despite the advances of the last three decades, speech recognition research is still far from achieving its ultimate goal—namely, automatic recognition by machine of unconstrained speech from any speaker in any environment. There are several key reasons why this remains a challenging problem. One important factor is the difficulty of choosing the set of speech units used for recognition. The optimal choice seems to depend upon the task. Units as large as words and phrases are ideal for limited task environments (e.g., dialing telephone numbers, voice command words for text processing); however, these large speech units are totally unsuitable for speech recognition involving large vocabularies and/or continuous, naturally spoken sentences. There are two main reasons for this. First, a prohibitively large amount of data needs to be collected and labeled for establishing the necessary statistical knowledge. (Consider the extreme case of using sentences as the units in continuous speech recognition. The number of possible sentences becomes astronomical even with a medium vocabulary size. Collecting all the realizations of these sentences is infeasible if not outright impossible.) Second, even if such a large amount of data were to be collected, the complexity of the search during the recognition phase would be prohibitive. For such applications, sub-word recognition units (e.g., dyad, diphone, syllable, fractional syllable, or even phoneme) become necessary, and techniques for composing words from such sub-word units (such as lexical access from stored pronouncing dictionaries) are needed. Since such units are often not well articulated in natural continuous speech, they can be decoded only with a high error rate. Grouping these error-prone sub-word sequences into sequences of words from a standard lexicon becomes a very difficult task.

Another important factor affecting performance is the size of the user population. So-called speaker-trained or speaker-dependent systems adapt to the voice patterns of an individual user via an enrollment procedure. For a limited number of frequent users of a particular speech recognizer, such an enrollment procedure works reasonably well and generally leads to good recognizer performance. However, for applications where the user population is large and the usage casual (e.g., users of automatic number dialers or order-entry services), it is infeasible to train the system on individual users. In such cases, the recognizer must be speaker independent and able to adapt to a broad range of accents and voice characteristics.

Other factors that affect recognizer performance include:

- Complexity of the vocabulary
- Transmission medium over which recognition is performed (e.g., over a telephone line, in an airplane cockpit, or using a hands-free speakerphone)
- Task limitations in the form of syntactic and semantic constraints on what can be spoken
- Cost and method of implementation

This partial list of factors illustrates why unconstrained automatic speech recognition by machine is unlikely to be achieved in the foreseeable future. Much rigorous research remains to be done.

Currently, speech recognition has achieved modest success by limiting the scope of its applications. Considerable success has been achieved with systems designed to recognize small- to moderate-size vocabularies (10 to 500 words) in a speaker-trained manner in a controlled environment with a well-defined task (e.g., order entry,

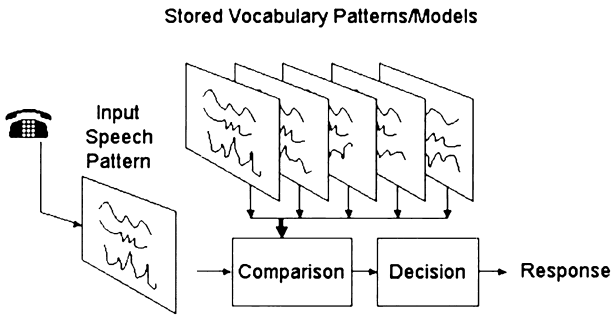


FIGURE 5 Automatic speech recognition using a pattern recognition framework.

voice editing, telephone number dialing) and spoken cooperatively (as opposed to conversational utterances full of disfluencies such as partial words, incomplete sentences, and extraneous sounds like "uh" and "um"). Such systems have been, for the most part, based entirely on the techniques of statistical pattern recognition, as discussed above. In the simplest context, each word in the vocabulary is represented as a distinct pattern (or set of patterns) in the recognizer memory as shown in Fig. 5. (The pattern can be either a sample of the vocabulary word, stored as a temporal sequence of spectral vectors, or a statistical model of the spectrum of the word as a function of time.) Each time a word or sequence of words (either isolated or connected) is spoken, a match between the unknown pattern and each of the stored word vocabulary patterns is made, and the best match is used as the recognized string. In order to do the matching between the unknown speech pattern and the stored vocabulary patterns, a time-alignment procedure is required to register the unknown and reference patterns properly because speaking rate can vary between wide limits. Several algorithms based on the techniques of dynamic programming have been devised for optimally performing the match.

Tables I and II provide a summary of typical speech recognition performance for systems of the type shown in Fig. 5. Table I shows average word error rate for context-free recognition of isolated words for both speaker-dependent (SD) and speaker-independent (SI) systems. (By context-free, we mean that no task syntax or semantics is available to help detect and correct errors.) It can be seen that, for the same vocabulary, SD and SI recognizers can achieve comparable performance. It is further seen that performance is more sensitive to vocabulary complexity than to vocabulary size. Thus, the 39-word alpha-digits vocabulary (letters A–Z, digits 0–9, three command words) with highly confusable word subsets—such as B, C, D, E, G, P, T, V, Z, and 3—has an average error rate of about 5 to 7%, whereas an 1109-word basic English vocabulary has an average error rate of ∼4%.

TABLE I Performance of Isolated Word Recognition Systems

Task/application	Vocabulary size	Mode	Word accuracy
Digits	10	SI	∼100%
Voice dialer words	37	SD	100%
Alpha-digits plus command words	39	SD	96%
		SI	93%
Computer terms	54	SI	96%
Airline Words	129	SD	99%
		SI	97%
Japanese city names	200	SD	97%
Basic English	1109	SD	96%

Table II shows the typical performance of connected-word recognizers applied to tasks with various types of constraints. The highly accurate performance on connected digits (both speaker-trained and speaker-independent) is a result of significant advances in the training procedures for recognition. In the table, perplexity is defined as the average number of words that can follow an arbitrary word. It is usually substantially less than the size of the vocabulary because of the syntactic, grammatical, and semantic constraints of the task-specific language. It is a rough measure of the difficulty of the task and may be estimated from a statistical model of the language that defines the sequential dependence between words. For example, an N-gram model defines the probability of a sequence of N words (or units) as observed in a large collection of sentences in the language. Language models are obtained essentially by word counting procedures.

The Naval Resource Management task involves a particular kind of language used in naval duties and is highly stylized. Sentences in the Air Travel Information System are query utterances focusing on flight information such as flight time, fare, and the origin and destination of the intended travel. The vocabulary size reflects the number of city or airport names included in the database. The grammatical structure in the query sentence is quite limited compared to many other application domain languages. In the *Wall Street Journal* transcription task, the input is "read" speech recorded with a reasonably good-quality microphone. Read speech is known to be much easier to recognize than spontaneous, conversational speech. The Broadcast News transcription task involves signals often referred to as "found" speech such as radio or television news announcements. The signal may be degraded due to noise and distortion, and the spontaneity is generally somewhat higher than for "read" speech. The degradation in recognition accuracy is a clear indication that this is a more difficult task. These tasks, however, are all considered far easier than transcribing and understanding truly spontaneous conversational speech.

TABLE II Performance Benchmark of HMM-Based Automatic Speech Recognition Systems for Various Connected and Continuous Speech Recognition Tasks or Applications

Task/application	Vocabulary size	Word Perplexity	Accuracy
Connected digit strings	10	10	~100%
Naval Resource Management	991	<60	97%
Air Travel Information System	1,800	<25	97%
Wall Street Journal transcription	64,000	<140	94%
Broadcast News transcription	64,000	<140	86%

The systems listed in Table II use phoneme-like statistical unit models to represent words, as do many speech recognition software packages offered on the market for personal computing (PC) applications such as dictating a letter. These phoneme-like unit models are also modified by the context in which they appear. For example, the /l/ in /e-l-i/ is different from the /l/ in /o-l-o/. Many systems employ thousands of such context-dependent unit models in continuous speech recognition. The benchmark results in Table II are based on very extensive speaker-independent training. On the other hand, PC-based software systems normally require speaker adaptation. The user of PC-based speech recognition software is asked to speak a designated set of sentences, ranging from 5 minutes to a few hours in duration, so that parameters of the baseline system can be modified for improved performance for the specific user.

As illustrated above, the performance of current systems is barely acceptable for large vocabulary systems, even with isolated word inputs, speaker training, and favorable talking environment. Almost every aspect of continuous speech recognition, from training to systems implementation, represents a challenge in performance, reliability, and robustness.

Another approach to machine recognition and understanding of speech, particularly for automated services in a limited domain, is the technique of keyword spotting. A word-spotting system aims at identifying a keyword (or a key phrase), which may be embedded in a naturally spoken sentence. This is very useful because it makes the interaction between the user and the machine more natural and more robust than with a rigid command-word recognition system. Experience shows that many (infrequent) users of a speech recognition system in telecommunication applications often speak words or phrases that are not part of the recognition vocabulary, thus creating so-called out-of-vocabulary (OOV) or out-of-grammar (OOG) errors. Rather than attempting to recognize every word in the utterance, a word-spotting system hypothesizes the presence of a keyword in appropriate portions of the speech utterance and verifies the hypothesis by computing two scores, one for the hypothesized portion of the speech signal and the keyword model, and the other for the former and a background speech model. These scores are subject to a ratio test against a threshold for the final decision. By avoiding forced recognition decisions on the unrecognizable and inconsequential regions of the speech signal, the system can accommodate natural command sentences with good results as long as the number of keywords is limited (say, less than 20). Today, a word-spotting system with five key phrases (*collect*, *credit card*, *third party*, *person-to-person*, and *operator*) has been extensively deployed. By automating billions of the telephone calls traditionally categorized as operator-assisted calls, the system provides tremendous savings in operating cost.

VI. SPEAKER VERIFICATION

The objective of speaker verification is authentication of a claimed identity from measurements on the voice signal. Applications of speaker verification include entry control to restricted premises, access to privileged information, funds transfer, credit card authorization, voice banking, and similar transactions.

There are two types of voice authentication: one verifies a talker's identity based on talker-specific voice characteristics reflected in spoken words or sentences, and the other based on the content of the spoken *password* or *passphrase*, such as a personal identification number (PIN), social security number, or mother's maiden name. In the former case, the test phrase may be in the open and even shared by talkers in the population, while in the latter case the password information is assumed to be known only to the authorized talker. We often refer to the former as speaker verification (SV) and the latter as verbal information verification (VIV).

Research in speaker verification has a much longer history than VIV and in the past has encompassed studies of acoustic and linguistic features in the speech signal that carry the characteristics of the talker. In recent years, the dominance of a statistical, data-driven pattern matching approach (see Section V) has substantially changed the research landscape. The assumption is that a powerful statistical model such as the hidden Markov model would be able to automatically "discover" talker-specific characteristics provided that sufficient spoken utterances from the particular talker are available to allow such learning. Indeed, the statistical modeling technique has made a similar impact on speaker verification as on speech recognition. It is based on a principle almost identical to the one

illustrated in Fig. 5 in which the vocabulary models are replaced with the talker models or voice patterns. Each of the stored talker-specific models has to be trained according to an enrollment procedure before the system is put to use. In telephony applications these models may be stored in some remote location. For verification, the talker makes an identity claim (e.g., an account number) and speaks a test phrase (or simply the account number itself). The system compares the input speech with the stored model or pattern for the person with the claimed identity. On the basis of a similarity score and a carefully selected decision threshold, the system can accept or reject the speaker. The features useful for verification are those that distinguish talkers, independent of the spoken material. In contrast, the features useful for speech recognition are those that distinguish different words, independent of the talker. The decision threshold is often made dependent on the type of transaction that will occur as a result of the verification process. Clearly, a more stringent acceptance threshold is required for the transfer of money from one account to another than for reporting a current balance in a checking account. Key factors affecting the performance of speaker verification systems are the type of input string, the features that characterize the voice pattern, and the type of transmission system over which the verification system is used. Best performance is achieved when sentence-long utterances are used in a relatively noise-free speaking environment. A state-of-the-art system using a text-dependent test sentence is capable of achieving a 1 to 2% equal error rate (when the threshold is adjusted such that the probability of false acceptance and that of false rejection becomes equal). Conversely, poorer performance is achieved for short, unconstrained spoken utterances in a noisy environment (4 to 8% equal error rate using text-independent isolated words).

The challenge in speaker verification is to build adaptive talker models based on a small amount of training that perform well even for short input strings (e.g., one to four words). To achieve this goal, more research is needed in the area of talker modeling as well as in the area of robust analysis of noisy signals.

A VIV system stores the confidential information about a talker in a profile. When an identity claim is made, the talker is asked one or a series of questions based on the stored information; for example, "Please say your PIN" or "What is your birth date" or "What is your mother's maiden name?" The system verifies the information by scoring the spoken response with a composite speech model of the expected password or pass phrase constructed from the appropriate unit phone models. Such a system can achieve perfect (0 error) verification after three rounds of questions and answers.

VII. SPEECH ENHANCEMENT

During transmission from talker to listener, a speech signal may be degraded in a variety of ways. In this section, we will discuss some of these degradations and some of the methods that have been devised to deal with them.

Most speech enhancement algorithms are concerned with minimizing the perceptual effects of the degradation that results from additive noise. Noise can get mixed into a speech signal in several ways. For instance, speech may originate in a noisy environment, such as in a noisy airport, railway station, or shopping mall. A microphone placed in such an environment will pick up the sum of the desired speech and the ambient noise. Examples of practical devices that may be subjected to such degradation include cellular handsets, speakerphones, etc. Even low levels of ambient noise can become a problem in multipoint teleconferencing. This is because each participant in such a teleconference hears the ambient noise from all remote locations. Thus, assuming similar conditions at all locations, the noise would be about 10 dB higher in a teleconference among ten participants than in a two-way conference. Noise can also enter a conversation electrically in transmission lines, filters, amplifiers, etc., that are part of a telephone circuit.

In most such situations, all that is available to a speech enhancement algorithm is a single microphone signal representing the sum of the desired speech signal and the noise. Hence, to improve the quality of such degraded speech, the enhancement algorithm must operate "blindly." That is, it must operate in the absence of any prior knowledge of the properties of the interfering noise. It must estimate the noise component from the noisy signal and then attempt to minimize the perceptual effects of the noise.

There are, however, important applications in which additional information is available, besides the noisy speech signal. A prime example of such applications, and one that has been of interest for several decades, is echo cancellation on long-distance telephone circuits. Due to impedance mismatches on such circuits, an undesirable echo of a speech signal gets added to the desired signal being transmitted. This echo is the "noise" that needs to be eliminated from the "noisy" signal. If the speech signal that is responsible for the echo is observable, then the echo can be estimated and subtracted from the noisy signal. Another example of such applications might occur when a microphone is used to record the speech of a talker in a room with a noise source—say, an air-conditioning duct. If it is possible to place a secondary microphone very close to the noise source, then its output can be used to estimate and cancel the noise present at the recording microphone.

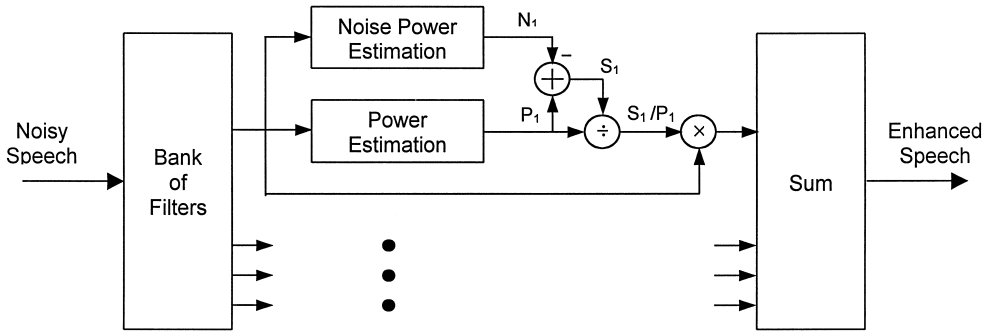


FIGURE 6 A speech enhancement system based on spectral subtraction.

A third type of speech enhancement is required to overcome the loss of intelligibility due to hearing loss that, in general, increases with age. Hearing loss is becoming more and more important in telephony, as life-expectancy increases and the percentage of older people in the population increases.

Finally, speech enhancement is of interest in situations where the speech may originate in a reasonably quiet environment but is to be received by a listener who is in a noisy environment. An example of such a situation is announcements over the public address (PA) system at a railway platform. In such cases, the noise itself is not under the control of a speech enhancer. However, if some estimate of the noise is available, then it is possible to preprocess the speech signal so as to improve its intelligibility as well as its spectral balance when listened to in the noisy environment.

Let us take a closer look at each of these four applications of speech enhancement.

In the case when only the noisy speech signal is available, speech enhancement is accomplished by systems that may be typified by the block diagram shown in Fig. 6. The noisy speech is passed through a bank of contiguous band-pass filters that span the frequency range of the speech signal (say, 200 to 3500 Hz for telephone speech). At the output of the ith filter, the short-term power, P_i, is an estimate of the power spectrum of the noisy speech at the center frequency of the filter. If the desired speech signal and the noise are uncorrelated, then $P_i = S_i + N_i$, where S_i is the power of the speech signal and N_i is the power of the noise at that frequency. Suppose for a moment that the noise power, N_i, can be estimated independently. (One way to estimate the noise power is to detect time intervals during which speech is present and measure the output power only during the other intervals. Several algorithms are known for detecting the presence of speech.) Then, $S_i = P_i - N_i$ is an estimate of the signal power. The ith channel signal is next adjusted by multiplying it with a gain factor S_i/P_i. Finally, these adjusted signals are all added together to yield the enhanced speech. It is seen that the output of a channel is attenuated more if it has more noise in it. This is what provides the enhancement.

Instead of introducing time-varying gains in the channels of a filter bank as shown in Fig. 6, an alternative approach derives the control signals of the type shown in Fig. 1 from the given noisy speech and then resynthesizes speech from these control signals. The first step as before is to estimate the signal power, S_i, in each channel and from it the signal amplitude (which is just the square root of the signal power). These estimated amplitudes characterize the time-varying filter of Fig. 1. In a parallel path, the given noisy signal is analyzed to determine intervals in which the signal is voiced and the intervals in which it is unvoiced. During voiced intervals, the fundamental frequency is determined. It turns out that such analysis can be performed reliably even in the presence of noise which is only 6 to 10 dB below the level of the speech signal. In this manner, the characteristics of the excitation source of Fig. 1 are determined. The enhanced signal is now synthesized as in Fig. 1.

All single-microphone enhancement systems in use today are variations of one of the two methods described above. They differ in the detailed characteristics of the band-pass filters, the method of estimating the noise power, the method of determining the fundamental frequency, etc.

These speech enhancement methods reduce the level of noise but unfortunately also introduce artifacts of their own. The main artifact is what is known as *musical noise*, which consists of bursts of periodic signals with randomly varying frequency. Many people prefer the original noisy signal to the processed signal because the musical noise is quite disconcerting to listen to. However, this type of artifact is almost inaudible when the noise in the original signal is at least 10 dB or more below the signal level. Also, enhancement system parameters can be adjusted so as to trade off this type of musical noise distortion against the residual noise. It should also be noted that except in

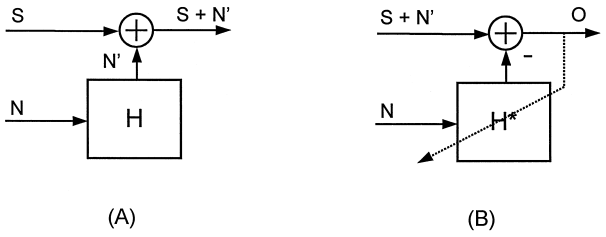


FIGURE 7 (A) Model for signal contamination; (B) Setup for canceling contamination from the desired signal.

one study with specialized test signals, these methods have *not* been shown to increase intelligibility. The main advantage of these speech enhancement methods is that they reduce fatigue that results from prolonged listening to noisy speech.

Algorithms for the second type of enhancement—echo cancellation or noise cancellation—are illustrated in the block diagram shown in Fig. 7. In Part A of this figure, the signal (S) is the desired signal to be recorded or transmitted. To that signal, an undesired signal (N') gets added. The signal N' is not known; however, it is known that N' is the output of a (hitherto unknown) linear filter (H), whose input is a known signal (N). In the echo cancellation example mentioned earlier, N' would be the echo generated by the signal N, and H would be the filter representing the transfer function of the echo path. In the other example, N' would be the noise signal at the recording microphone due to the noise N at the air-conditioning duct as picked up by the secondary microphone, and H would be the acoustic transfer function of the room from the air-conditioning duct to the recording microphone. An enhancement system for the echo or noise cancellation system is shown in part B of Fig. 7. The box marked H^* denotes an estimate of the filter (H). If we could make H^* equal to H, then clearly the noise would be removed and the output signal (O) would be the same as the desired noise-free signal (S). To drive the transfer function of the filter H^* towards that of H, an iterative gradient algorithm is used. During silent intervals of the speech signal (S), the algorithm adjusts H^* in such a way as to drive output O towards zero. If O can be made exactly zero in those intervals, then H^* provides a good estimate of H. If now the adaptive algorithm is turned off whenever the speech signal is present, then the noise portion of the signal, O, continues to be canceled, hence O approximates the desired noise-free signal S.

Such algorithms have been highly successful in the echo cancellation application. Indeed, several million such devices for echo cancellation have been deployed in various telephone systems. Their usage for the other application—reduction of room noise—is far less. First of all, there are far fewer occasions where this type of noise reduction is needed. Besides that, except at low frequencies, the noise at the recording microphone in a room cannot be accurately modeled as a single signal passed through a linear filter. In most cases, the noise at the microphone either lacks coherence with the noise reference (recorded at another location) or results from multiple, diffuse noise sources which cannot be represented by a single reference signal. Only in some special cases, e.g., in an airplane cockpit, or in a racing car, communication from the pilot or the driver can be somewhat improved by this technique.

Let us now turn to the third source of degradation mentioned previously, i.e., the degradation due to hearing loss. Fig. 8(a) shows the threshold of hearing as a function of frequency for a person with normal hearing. Also shown in that figure is the threshold of discomfort as a function of frequency. Any sound louder than this threshold would be

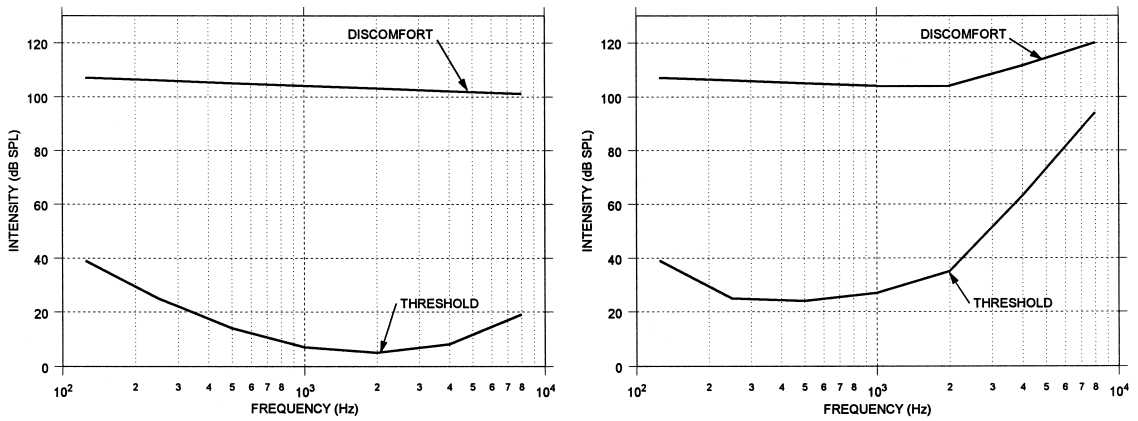


FIGURE 8 Typical thresholds of hearing and discomfort: (a) normal ear, (b) ear with hearing loss.

painful and can be harmful to the ear. Figure 8(b) shows the same two curves for a person with hearing loss. This figure is for illustration purposes only; there are many types of hearing loss with different characteristics.

The main noteworthy feature common to all types of hearing loss is that the threshold of hearing becomes higher than that for a normal ear, but the threshold of discomfort remains more or less unchanged. Also, in general at high enough intensities, the loudness at any frequency is the same for a person with normal hearing as for one with hearing loss. This last phenomenon is known as *loudness recruitment*.

A vast majority of hearing aids in use today provide only linear amplification, with frequency-dependent gain. From Figs. 8a and b, it is clear that this is not a satisfactory solution. If the gain is adjusted to be correct for a low-level sound, it will be too high for a high level sound, and vice versa. What is needed is to compress the dynamic range of sound at any frequency to fit the reduced dynamic range due to the hearing loss. Thus, an amplifier is needed such that its gain depends not only on the frequency but also on the signal level at that frequency so as to provide the appropriate compression. Hearing aids are now available that filter the speech signal into a number of bands (usually limited to two or three due to computational complexity) and provide a compression in each band that is appropriate for the hearing loss being corrected. This technique is called *multiband compression*.

Much improvement is still possible in this type of hearing aids. We do not yet know the optimal control strategy for the signal-dependent gain nor do we know the best procedures for fitting such a hearing aid. The fundamental reason for this is that the information summarized in Figs. 8a and b is obtained from tests with stationary signals and is inaccurate for a signal such as speech, which is continually changing. Also, note that the high gain required for low-level sounds also amplifies the noise. For this reason, hearing aids tend to be unsatisfactory in noisy environments. In the future, with better understanding of the hearing process, as well as with the availability of smaller and faster digital processors with low power requirements, hearing aids that combine noise reduction with signal-dependent gain will no doubt be available and will provide much better solutions to the problem of hearing loss.

Finally, let us consider the preprocessing of a speech signal to make it more intelligible in the presence of ambient noise. It turns out that the noise induces something akin to hearing loss in a normal hearing person. At any frequency, the threshold of hearing is elevated while a signal sufficiently above the noise level sounds almost as loud as in the absence of noise. Of course, the amount of loss depends on the level of noise at each frequency. Thus, if the clean speech signal is amplified with a gain that has the appropriate dependence on frequency and intensity, it can be made much more intelligible when listened to in the presence of noise. Of course, the speech will still be quite noisy, since the noise is not changed.

VIII. CONCLUDING REMARKS

This cursory overview of digital speech processing has been intended to highlight recent advances, current areas of research, and key issues for which new fundamental understanding is needed. Future progress in speech processing will surely be linked closely with advances in computation, microelectronics, and algorithm design.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICIAL INTELLIGENCE • COGNITIVE SCIENCE • DIGITAL CONTROLLERS • DIGITAL FILTERS • HUMAN-COMPUTER INTERACTION • INFORMATION THEORY • SOFTWARE ENGINEERING • SPEECH SYNTHESIS BASED ON LINEAR PREDICTION

BIBLIOGRAPHY

Flanagan, J. L. (1972). "Speech Analysis, Synthesis, and Perception," Springer-Verlag, New York.

Furui, S., and Sondhi, M. M., eds. (1991). "Advances in Speech Signal Processing," Dekker, New York.

Jayant, N. S., and Noll, P. (1984). "Digital Coding of Waveforms," Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L. R., and Gold, B. (1975). "Theory and Application of Digital Signal Processing," Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L. R., and Juang, B. H. (1993). "Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L. R., and Schafer, R. W. (1978). "Digital Processing of Speech Signals," Prentice-Hall, Englewood Cliffs, NJ.



Magnetic Recording

Kelly Knudson Fitzpatrick
Maxtor Corporation

I. Magnetic Recording Basics
II. Recording Technology
III. Signal Processing for Digital Recording

GLOSSARY

AC bias recording An analog recording technique where a high-frequency alternating current (AC) signal is added to the analog signal before recording.
Areal density The number of bits per unit area on the surface of the media.
Channel bits The binary sequence after encoding that corresponds to the magnetization pattern on the media.
Code rate The number of bits before encoding divided by the number of bits after encoding.
Coercivity The magnetic field required to switch the magnetization of the media from saturation in one direction to saturation in the opposite direction.
Error-correction coding (ECC) A method of embedding user bits in a codeword with a mathematical structure so errors can be corrected.
Frequency modulation (FM) recording An analog recording technique where the frequency of a high-frequency carrier signal depends linearly on the analog signal.
Head A device that writes or reads the magnetization pattern on the media.
Linear density The number of bits per unit length along a track.
Media The recording surface that consists of a magnetic layer supported on a nonmagnetic substrate.
Modulation code A mapping from one binary sequence to a longer binary sequence that satisfies a set of constraints.
Saturation The maximum magnetization achieved when all the magnetic particles in a region align in one direction.
Saturation recording A digital recording process where the magnetization pattern switches back and forth between saturation in one direction and saturation in the opposite direction.
Superparamagnetism Random magnetic fluctuations due to thermal energy that can cause a magnetization pattern to degrade over time.
Track A line of data stored on a magnetic tape, or a concentric ring of data stored on a magnetic disk.
Transition A region on the media where the magnetization switches from saturation in one direction to saturation in the opposite direction.
User bits Binary data bits transferred from the host computer.

MAGNETIC RECORDING is the process of storing information in the form of a magnetic pattern and retrieving that information at a later time. The recorded signal can be stored indefinitely, provided the magnetic medium

is not exposed to magnetic fields comparable in strength to those used in recording. Magnetic heads are used to record magnetization patterns along tracks on the media during writing and to generate playback signals during reading.

The field of magnetic recording is vast with major applications in audio, video, and data storage. Audio and video tape recorders have historically used analog recording, whereas computer-generated data are stored on disk drives and tape drives using digital recording. Since digital recording has the advantage that the original message can be reconstructed perfectly through the use of error-correction coding, the audio and video industries are moving towards digital recording of analog signals that have been sampled and converted to digital.

There are three general types of magnetic media: tape, hard disk, and flexible disk. Tape is used in all the major applications of audio, video, and data storage. Tape drives provide access to large quantities of information stored on inexpensive tape cassettes. The major drawback to tape is that random access is slow due to the mechanical process of rewinding the tape.

Hard-disk drives provide rapid access to data by positioning a head above a spinning disk. Spinning the disk faster reduces the seek time that elapses before a specific piece of data passes the head. The average seek time is only a few milliseconds. Currently, hard-disk drives for personal computers spin between 4500 and 7200 rotations per minute (rpm). For high-end server applications, the spin rate is as high as 10,000 rpm. The phenomenal growth in the computer industry has fueled an increase in disk-drive shipments of 25% per year over the last 40 years. In 1998, there were approximately 98 million 3.5-inch disk-drives shipped.

Flexible (floppy) disks enable data to be transferred from one computer to the next while still maintaining rapid access and relatively high capacity. Even though most personal computers still have flexible-disk drives, their role has diminished now that most computers are network connected.

Throughout the history of magnetic recording there has been a constant effort to increase the density of information stored on the surface of the magnetic media. The number of bits per square inch on the media surface (areal density) rises when either the number of tracks per inch (track density) or the number of bits per inch along a track (linear density) increases. Improvements in track density require more accurate positioning of the head, and improvements in linear density require more complex signal processing.

As the surface area devoted to each bit shrinks, the signal amplitude decreases dramatically. Technological advancements in heads and media are required to increase head sensitivity and decrease noise to maintain an adequate signal-to-noise ratio (SNR).

Over the last 40 years, areal densities on hard-disk drives have doubled every 18 months and at times have doubled every 12 months. This tremendous growth in capacity has been achieved while simultaneously decreasing the cost per drive.

An overview of the basic principals of analog and digital magnetic recording is provided in Section I. Section II takes a look at the technological developments in magnetic heads and media that have enabled advancements in areal densities. Since digital magnetic recording is such a major aspect of today's computerized world, Section III describes in detail the signal processing performed in the disk drives and tape drives used for mass storage.

I. MAGNETIC RECORDING BASICS

In general, a magnetic recording device is part of a larger audio, video, or computer system. During writing, the signal from the host is transformed into a write current signal that is sent to the write head and recorded as a magnetic pattern. During reading, a read head senses the magnetic pattern and produces a readback signal that is processed and sent back to the host.

The writing process involves the relative motion between the write head and the media along a track. The write current signal causes the head to generate an applied magnetic field that magnetizes the media in the proximity of the head. The magnetization pattern left on the media is not a simple linear function of the applied field. In fact, the relationship depends on the direction and magnitude of previously applied fields and is best described graphically by a hysteresis loop (Fig. 1). The tiny magnetic particles in the medium rotate in response to an external field. Once all the magnetic particles are aligned, continuing

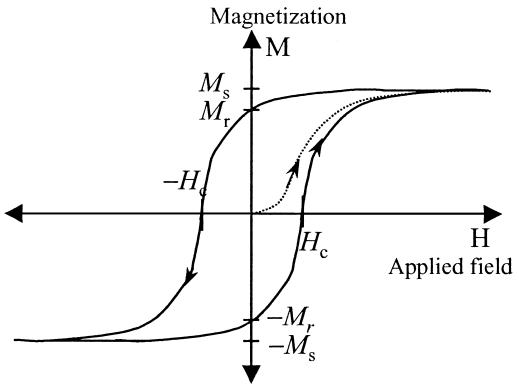


FIGURE 1 Hysteresis loop showing the nonlinear relationship between magnetization *M* and an applied field *H*.

to increase the applied field does not increase the magnetization. The magnetization saturates at the saturation level, M_s. When the applied field is removed, the particles relax to a remanent magnetization, M_r, where most of the particles remain aligned. The magnetic media remembers the applied field, hence the name remanent. Magnetic recording is possible because magnetic media have memory.

When transitioning from the negative saturation level to the positive one, the magnetization goes to zero when the applied field is equal to the coercivity of the media, H_c. However, when transitioning from the positive to the negative saturation level, the magnetization takes an entirely different path that goes to zero when the applied field is $-H_c$.

The coercivity expresses the degree to which the magnetized media resists demagnetization. The strength of the magnetic medium is measured by the coercivity times the remanent magnetization, $M_r H_c$.

A. Digital Recording

Digital recording is the process of storing binary data in the form of a two-level magnetization pattern. Digital recording is also referred to as saturation recording since channel bits are recorded by saturating the media either in the positive or negative direction. The resulting magnetic pattern alternates between the two remanent magnetization levels, $+M_r$ and $-M_r$.

During writing, binary channel bits with values 0 and 1 are converted to square pulses with duration T and amplitudes -1 and 1. These square pulses make up a two-level write current signal that is sent to the write head. The resulting magnetization pattern is a track of contiguous bar magnets with magnetization M_r or $-M_r$, depending on the channel bit values. The physical length of a channel bit is $L = vT$, where v is the relative velocity between the write head and the media.

When the input to the digital recording channel is a unit pulse, with amplitude 1 and duration T, the output is called the pulse response. A pulse response has a negative peak and a positive peak and looks something like Fig. 2.

During reading, the readback signal is a linear combination of pulse responses staggered in time and scaled by $+1$ or -1. Figure 3 shows an example of the two-level write current signal, the magnetization pattern, and the readback signal for a typical channel bit sequence. The dashed lines in the figure show the individual pulse responses that are summed together to form the readback signal.



FIGURE 2 An example pulse response for a magnetic recording channel.

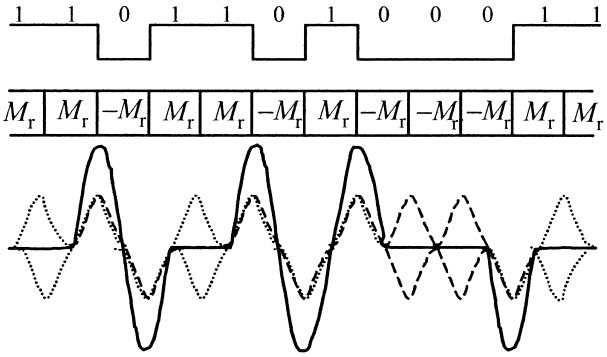


FIGURE 3 From top to bottom: binary channel bits, two-level write current signal, magnetization pattern, and readback signal.

One of the major benefits of using digital recording is the ability to use error-correction coding to improve the integrity of the recorded data. Error-correction coding encases the data in a mathematical superstructure, so that if some of the data are detected incorrectly, the overriding structure is still evident and the data can be corrected. Signal processing and coding for digital magnetic recording are explored further in Section III.

B. Analog Recording

In analog recording, the magnetization pattern is written in such a way that the readback signal is proportional to an analog input signal. There are two basic analog recording methods: AC bias and frequency modulation. AC bias is the standard analog recording method used in audio tape recorders. Frequency modulation is used for video recording of analog television signals. The two techniques differ since audio signals have lower bandwidth than television signals.

1. Audio Tape Recording

In audio tape recorders, a special technique is used to make the magnetization a linear function of the applied field. This is a challenging feat since the relationship between magnetization level and applied field is highly nonlinear, as shown by the hysteresis curve in Fig. 1. The recording process is made linear by adding a high-frequency alternating current (AC) signal to the audio signal. The frequency and amplitude of the bias signal must be many times greater than the largest frequency and amplitude of the audio signal. Under these conditions, the hysteretic

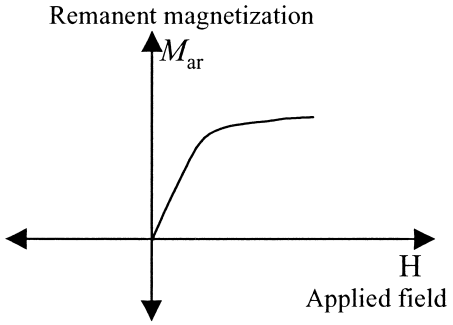


FIGURE 4 Remanent magnetization M_{ar} as a function of applied field H.

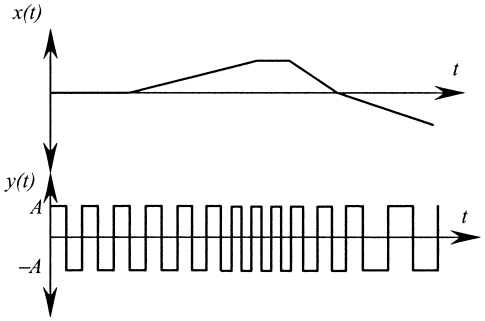


FIGURE 5 A hard-limited FM modulated signal $y(t)$ corresponding to an analog input signal $x(t)$.

effects are averaged out to produce an anhysteretic relationship between remanent magnetization and applied field, as shown in Fig. 4. The remanent magnetization is the magnetization level left after the applied field is removed.

Attenuation at high frequencies is the principal audio signal impairment. Ideally, the frequency response should be flat over the entire audio bandwidth. Audio recorders use a combination of pre-equalization and post-equalization to compensate for high-frequency losses. Pre-equalization is linear filtering performed on the audio signal prior to recording, in order to boost high frequencies. Post-equalization is linear filtering performed after reading back the signal. Post-equalization increases the noise, whereas pre-equalization does not. Since audio tapes recorded on one machine must be playable on any other machine, the post-equalization is standardized. As a consequence, the pre-equalization is adjusted to produce a flat response for the particular head–media combination used for recording.

Several signal processing techniques have been developed for improving the SNR. Most notably, Dolby introduced a method for boosting low-level, high-frequency signals during recording and attenuating them by the same amount on reproduction (Dolby, 1967). During recording, low-amplitude audio signals are amplified to bring the remanent magnetization well above the noise level. During playback, the Dolby system senses when to switch the playback amplifier to a lower gain setting. In addition, distortion at high frequencies is reduced by dynamically reducing the amplitude of the AC bias current when the high-frequency energy in the audio signal is large.

To avoid nonlinear distortion due to saturation of large-amplitude audio signals, the dynamic range of the audio signal must be scaled to fit the linear portion of the curve in Fig. 4. This limitation on dynamic range corresponds to a loss in SNR on the order of 6 dB relative to saturation recording (Mintzer and Howell, 1982).

2. Video Tape Recording

In video tape recording, the analog television signal is frequency modulated during the recording process. In frequency modulation (FM), the frequency of a high-frequency carrier signal depends linearly on the analog input signal $x(t)$. When the video signal is zero voltage (DC), the frequency of the FM modulated signal is constant and equal to the carrier frequency, f_c. When the voltage of the video signal is positive or negative, the FM modulator changes the frequency up or down, respectively. The FM modulated signal is given by:

$$y(t) = A\cos\left(2\pi f_c t + k_f \int^t x(\alpha)\,d\alpha + \phi_0\right)$$

where A is the carrier amplitude, k_f is the frequency deviation constant, and ϕ_0 is the initial phase offset. During the recording process, the FM signal is essentially passed through a hard limiter so that the magnetization pattern recorded is actually a square wave signal rather than a sinusoid, as shown in Fig. 5. The resulting magnetization pattern carries frequency information in the spacing between transitions on the media.

The head used for sensing the recorded pattern differentiates and low-pass filters the signal. The low-pass filtering removes the high-frequency content, leaving a readback signal that approximates the analog video signal.

II. RECORDING TECHNOLOGY

Advancements in head and media technology have enabled tremendous growth in areal density, mainly by introducing new materials for smaller heads and thinner media. This section describes briefly the different types of magnetic media and the two basic types of heads: inductive and magnetoresistive. The signal and noise properties of heads and media are not the only consideration. These components must be designed to maintain recording quality over

a wide range of temperature and humidity as well as over a lifetime of extensive wear.

A. Magnetic Media

The various applications of magnetic recording generally use different types of media. Since each application has different linear density, mechanical stresses, and pricing tradeoffs, the optimal solutions vary considerably.

Because the signal amplitude falls off exponentially with the distance of the head from the media, improvements in areal density have been made by decreasing the thickness of the media and by lowering the fly height of the head. The second option only applies to a hard disk, since the head is already in contact with tape and flexible disk.

In general, the magnetic medium is composed of a uniform magnetic layer on top of a nonmagnetic substrate. For tape and flexible disks, the substrate is a polyester or Mylar film. The substrate is painted with a coating of magnetic particles, typically gamma ferric oxide, mixed in a binding solution. For tape, the orientation of the particles is often controlled during the coating process by applying a strong magnetic field.

For hard disks, the substrate is made of aluminum, glass, or ceramic. A very thin magnetic film is sputtered onto the disk substrate. The densely packed crystalline structure of thin-film media is required to achieve higher areal densities. Since these deposited films have higher magnetization than particulate coatings, a thinner layer can produce the required signal amplitude. Thin-film materials are usually cobalt-based alloys that possess a hexagonal, rod-shaped structure.

The axis for easiest rotation of a rod-shaped magnetic particle is in the direction of its longest dimension. For longitudinal recording, the particles have this easy axis along the track parallel to the media surface. For perpendicular recording, the easy axis is in the vertical direction, perpendicular to the media surface.

Longitudinal recording currently has widespread utilization, whereas perpendicular recording is an emerging technology. When a channel bit consists of only a few magnetic particles and thermal energy is sufficient to flip the magnetization of a particle, the random magnetic fluctuations due to thermal energy can cause data to disappear over time. The resulting thermal instability is referred to as superparamagnetism. Compared to longitudinal media, perpendicular media have more individual particles per unit surface area; therefore, the superparamagnetic limit occurs at a higher areal density for perpendicular recording. Hard-disk drives may need to use perpendicular recording in the near future to achieve the aggressive areal densities projected for future products.

B. Inductive Heads

Inductive heads are used for writing and reading. As a cost-savings measure, the same inductive head can be used for both writing and reading. However, this requires compromises in performance, since the optimal properties of write heads and read heads are different. The most common type of inductive head is a ring head wrapped with a wire coil. A ring head consists of a magnetic ring with a very small gap on the surface facing the media. During writing, current flowing through the coil induces a magnetic field that magnetizes the ring. As the magnetic field arcs across the gap, a small portion of the field fringes out toward the media. The fringing field has contour lines that are semicircles about the gap centerline, as shown in Fig. 6. Since only the fringing fields are used for magnetizing the media, the head must be kept in very close proximity to the media in order to produce a magnetic field contour line that exceeds the coercivity of the media.

As a point on the media passes the write head, it begins to experience a small magnetic field in the perpendicular direction. As the point approaches the gap, the amplitude grows and the direction rotates. Directly above the gap, the field has its maximum amplitude and its direction is entirely longitudinal. As the point moves away from the gap, the field rotates in the opposite perpendicular direction as the amplitude decreases to zero. Since the head field has its maximum amplitude in the longitudinal direction, ring heads are perfectly suited for longitudinal recording.

When reading with a ring head, flux from the media causes the magnetically permeable ring to magnetize in the direction of the flux. As the head passes over a transition, where the media has magnets of opposite polarity abutting one another, the magnetization of the ring flips direction. Thus, the coil wrapped around the ring is exposed to a change in magnetic flux, which in turn causes current to flow through the wire. The current passing through the inductive coil produces a voltage proportional to the

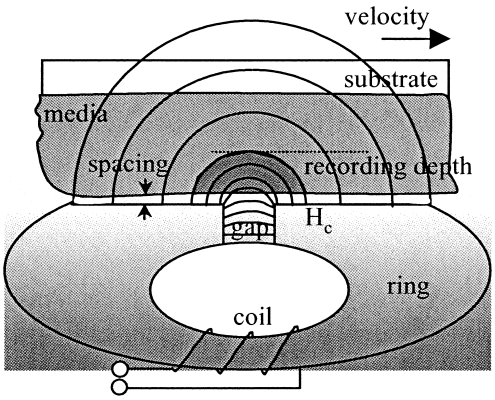


FIGURE 6 Magnetic field contour lines for a ring head.

number of turns the coil takes around the ring, N, and the rate of change of the current. Relating this back to the flux from the media, ϕ, one observes a playback voltage given by:

$$V(t) = -Nd\phi/dt$$

assuming there is contact between the ring and the media so that all the flux emanating from the media passes through the ring.

Since inductive heads sense the change in flux, the playback voltage is proportional to the relative velocity between the read head and the media. Let x denote the distance along the track and let $v = x/t$ represent the velocity, then

$$V(t) = -Nvd\phi/dx$$

Originally, inductive ring heads were fabricated from solid ferrite or laminated materials. The newer thin-film inductive heads have many advantages over the bulkier ferrite core heads. Their smaller size permits very high areal densities and their smaller inductance enables high-frequency operation without severe eddy current losses.

C. Magnetoresistive Heads

Magnetoresistive (MR) heads are only for reading; therefore, a separate inductive head is required to write the media. MR heads are often used in applications that do not require writing, such as credit card readers. Since magnetoresistive heads sense flux rather than change in flux, the voltage is independent of the velocity. This is an essential feature in hand-operated credit card readers.

Magnetoresistive elements are devices that change resistance in the presence of magnetic fields. By passing a current through the MR element, one can sense recorded signals on magnetic media. The signal-to-noise characteristics of magnetoresistive heads are far superior to those of thin-film inductive heads. In 1996, IBM started shipping disk drives with MR heads. Since then, most of the computer storage industry has joined the ranks.

Within a MR head there is a thin-film permalloy element, typically FeNi, whose electrical resistance changes a few percent when exposed to a magnetic field of a particular orientation. The amplitude of the readback signal is proportional to the change in resistance divided by the resistance.

New materials that exhibit a much larger magnetoresistive effect were introduced into commercial products in the year 2000. These elements are known as giant magnetoresistive (GMR) sensors. Even more sensitive MR sensors are under development and will likely displace GMR heads in the future.

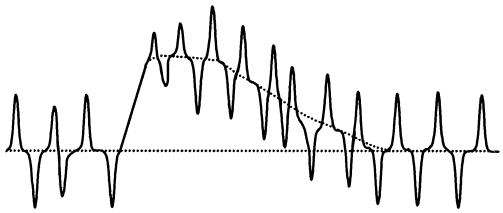


FIGURE 7 A thermal asperity while reading random data.

One problem that complicates the detection of playback signals occurs when a MR head is struck by a mechanical abrasion on the disk or tape. The contact between the two heats the MR element and causes a change in resistance. This is referred to as a thermal asperity. A thermal asperity causes a baseline shift in the playback signal that decays as the heat dissipates. Figure 7 shows a typical head response to a thermal asperity while reading random data. Since thermal asperities can cause a large burst of errors in the data, error-correction coding with large burst error-correction capability must be used to maintain data integrity.

D. Readback Signal Impairments

During reading, the head generates a relatively large voltage signal with a small time-varying component that is the playback signal of interest. An amplifier is used to remove the constant component and to magnify the tiny playback signal so that it is less vulnerable to electrical interference and has the voltage expected by the electronics. The three basic types of impairments in the readback signal at the output of the amplifier are noise, interference, and distortion. The ultimate goal is to minimize these impairments so that the signal-to-noise ratio is maximized.

The two major noise sources in recording systems are the electronics and the media. Electronic noise is caused by random fluctuations of electric charge carriers. In general, the electronic noise in the playback signal is spectrally flat and stationary. Media noise refers to fluctuations in the orientation and density of magnetic particles in the media. Ideally, all the magnetic particles should be uniformly magnetized in one particular direction. Variations in orientation and packing density cause perturbations in the playback signal. Media noise is written in during the recording process, therefore, the same media noise is experienced each time a particular bit is read. However, a different media noise pattern is written each time the track is rewritten. In contrast, electronic noise never repeats.

Interference refers to coherent signals that interfere with the intended signal. The principal source of interference is from data on adjacent tracks. Adjacent track interference

occurs when the head is positioned partially over the intended track and partially over an adjacent track during writing or reading. When the read head is narrower than the write head, the effects of track misregistration are mitigated at the expense of reduced on-track signal-to-noise ratio.

The last class of impairment in recording systems is nonlinear distortion. Nonlinear distortion is a deterministic impairment to the signal that occurs when the recording system is not linear. Under certain conditions magnetic recording systems satisfy the rules of linearity. For instance, a digital magnetic recording system is linear in the sense that the playback signal is a linear combination of pulses due to the individual input bits. When the input to a linear system is expressed as a linear combination of input signals, the output of the system is equal to a linear combination of the output signals associated with the individual input signals:

$$L[ax_1(t) + bx_2(t)] = aL[x_1(t)] + bL[x_2(t)]$$

Linearity breaks down when the transition spacing is made too small for the head and media used in a system. The demagnetization field from a closely spaced preceding transition causes the transition to shift with respect to where it would have been written in the absence of the preceding transition. Since the amount of unwanted shift is dependent on the particular bit pattern recorded, it is nonlinear.

Write precompensation is a commonly used technique for linearizing the channel in the presence of nonlinear transition shift. The transitions in the write current are shifted to precompensate for the demagnetization effect so that the transitions in the magnetization pattern occur in the appropriate positions.

Although magnetoresistive heads have good signal-to-noise characteristics, they also cause nonlinear distortion. The MR sensor saturates for large amplitude signals. Since the small amplitude signals are the dominant source of errors, the nonlinear distortion on the high-amplitude signal does not significantly effect performance.

III. SIGNAL PROCESSING FOR DIGITAL RECORDING

The signal processing performed during reading is an integral part of the data recovery process. As areal densities increase, advanced signal processing techniques are required to achieve reliable digital recording. A block diagram of the signal processing elements in a typical disk drive is shown in Fig. 8.

In the first stage of the write path, the user data from the host computer is encoded with an error-correction code (ECC). To further improve data integrity, the data sequence is modulation encoded before becoming the channel bits that are recorded in the form of a magnetic pattern. Modulation codes eliminate those binary sequences that most severely degrade the detector's performance.

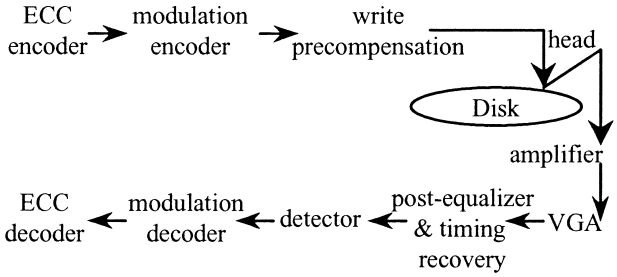


FIGURE 8 A block diagram of the signal processing for a disk drive.

During reading, the amplitude, timing, and shape of the signal must be controlled before entering the detector. This control is accomplished with a variable gain amplifier (VGA) that regulates the amplitude of the readback signal, a phase-lock loop that tracks changes in the phase and frequency, and a post-equalizer that shapes the signal.

In the final stages of the read path, the estimate of the channel bit sequence generated by the detector passes through a modulation decoder and an ECC decoder. As long as the number of errors does not exceed the error-correction capability, the original data are perfectly reconstructed.

The code rates of the modulation code and the ECC code are multiplied together to obtain an overall code rate. If we refer to the data bits transferred from the host computer as user bits, then the host computer is interested in storing a fixed number of user bits on the surface of the media. Given that the user bit density is fixed, a lower overall code rate allows for more error correction but reduces the size of the channel bits that are recorded on the media. If the increase in error correction is greater than the increase in errors due to smaller channel bits, then the code is said to have coding gain.

A. Peak Detection

Up until the 1990s all digital magnetic recording devices used peak detectors to convert the playback waveform to bits. At low recording densities, the transitions are spaced sufficiently far apart so that the response from one transition dies down before the next one starts. Under these conditions, the peaks in the playback signal are centered directly over the magnetic transitions in the media, as shown in Fig. 9. Peak detectors are well suited for detecting this type of low-density signal.

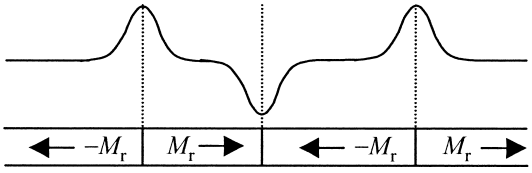


FIGURE 9 Peaks centered over transitions.

1. Modulation Coding for Peak Detection

Typically, a peak detector is used in combination with a modulation code. For simplicity, modulation codes are described using NRZI notation, where 1's represent transitions in the write current and 0's represent nontransitions. Previously, NRZ notation was used to represent the relationship between the channel bits and the write current. Fig. 10 gives an example of the NRZ and NRZI sequences associated with a write current signal.

Run length limited (RLL) modulation codes are used with peak detectors in recording systems. RLL codes are specified in terms of a d constraint that puts a lower limit on the number of 0's between two 1's, and a k constraint that puts an upper limit on the number of 0's between two 1's.

The d constraint is the most important parameter of a RLL code. It determines the size of a channel bit with respect to the minimum spacing between transitions, L_{tran}. Increasing the d constraint reduces the channel bit length L in relation to L_{tran}, since $L = L_{\text{tran}}/(d+1)$. If L_{tran} is fixed, then using a $(d=1)$ code doubles the channel bit density and increases the user bit density to twice the code rate.

Since some sequences are not allowed, the code rate must be less than 1. The capacity of a code constraint represents the maximum code rate of any code satisfying the code constraint. For a $(d=1, k=\infty)$ code constraint, the maximum code rate is 0.69. The most popular $d=1$ code is a rate 2/3 $(d=1, k=7)$ code that increases the user bit density by 32%.

A $(d=1)$ code does not contain the NRZI sequence {1 1} which corresponds to the channel bit sequences $\{+1\,-1\,+1\}$ and $\{-1\,+1\,-1\}$. A state transition diagram with two states, as shown in Fig. 11, describes this code constraint.

The $(d=1, k=7)$ code has a code rate of .66, which is close to the theoretical limit. The small k constraint ensures that transitions occur frequently. Since the phase-lock loop only updates its estimate of the phase and frequency when transitions are encountered, a small k constraint guarantees that there will be enough timing information to track changes in phase and frequency.

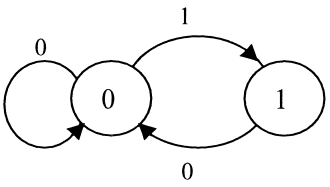


FIGURE 11 The state transition diagram for a $(d=1)$.

In a similar fashion, a rate 1/2 $(d=2, k=7)$ modulation code triples the channel bit density and increases the user bit density by 50%. These increases in user bit density are offset by a diminished SNR due to the increase in the bandwidth of the signal. Even though the signal and media noise are unchanged, the SNR decreases as the electronic noise scales inversely with code rate.

2. Peak Detection at Higher Recording Densities

At higher linear densities, the transitions are more closely spaced. The response from one transition interferes with the response from the next transition, such that the output of the channel is a linear combination of overlapping transition responses. Communication systems with this property are referred to as intersymbol interference channels.

Figure 12 shows what happens when a positive transition response is linearly superimposed with a closely spaced negative transition response. The destructive nature of the interference causes the magnitude of a crowded peak to be smaller than that of an isolated peak. In addition, the interference causes the position of the peak to shift away from the center of the transition in the magnetic media.

In order to use a peak detector on such a signal, steps must be taken to reduce the intersymbol interference at the peaks. During writing, the crowded transitions are written closer together than normal to precompensate for peak

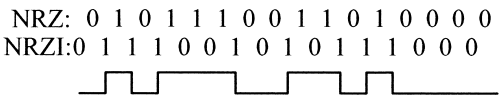


FIGURE 10 The NRZ and NRZI sequences associated with a write current signal.

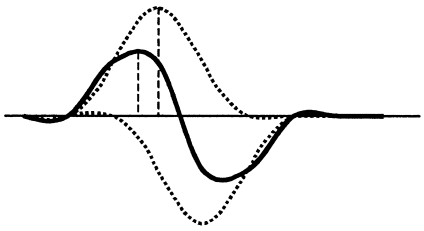


FIGURE 12 The superposition of two overlapping transition responses.

shift, so that the peaks are centered over the positions where the transitions would have normally been. In order to reduce the amplitude loss and peak shift, the signal is passed through a post-equalization filter that makes the pulses narrower by boosting high frequencies. While pulse slimming reduces interference, it increases electronic noise and therefore degrades detector performance when used too aggressively.

B. Partial Response Channels

When the intersymbol interference becomes significant, a more satisfactory solution is to allow it to occur, but to take the interference into account in the detection process. Instead of eliminating intersymbol interference, post-equalization is used to force the playback signal to have a controlled amount of intersymbol interference (ISI).

Bandlimited signals with controlled amounts of intersymbol interference are often referred to as partial response signals, which indicates that the response associated with a channel bit is only partially contained within one channel bit interval, T. Since the signal is bandlimited, the samples of a partial response signal taken every T seconds contain all the information required for detection. There is no information lost if the maximum frequency in the signal is less than $1/(2T)$. When this condition is not satisfied, there is some distortion at the high frequencies due to aliasing. Partial response signals are described by a discrete-time pulse response polynomial with coefficients $h_0, h_1, \ldots, h_m$ equal to the pulse response at the sampling times $0, T, 2T, \ldots, mT$:

$$h(D) = h_0 + h_1 D + \cdots + h_m D^m$$

where D is a unit-time delay operator. For those who are familiar with z-transforms used in digital signal processing, the D operator is the same as z^{-1}.

The samples of the partial response signal at the output of the post-equalizer are represented by the vector $\underline{r} = \{r_1, \ldots, r_n\}$. The sample at time iT is $r_i = y_i + n_i$, where y_i is the ideal noiseless part of the sample and n_i is the noise.

The ideal samples of a partial response signal can be constructed by convolving the input sequence $\underline{x} = \{\underline{x}_1, \ldots, x_n\}$, where $x_i \in \{-1, +1\}$, with the discrete-time pulse response:

$$y_i = \sum_{j=0}^{m} h_j x_{i-j}$$

The h_0 coefficient indicates how much the current input symbol x_i contributes to the current ideal sample value y_i. The other coefficients $\{h_m, h_{m-1}, \ldots, h_1\}$ measure the contribution from each of the m previous symbols $\{x_{i-m}, x_{i-m+1}, \ldots, x_{i-1}\}$, respectively. The contribution from these symbols is the intersymbol interference.

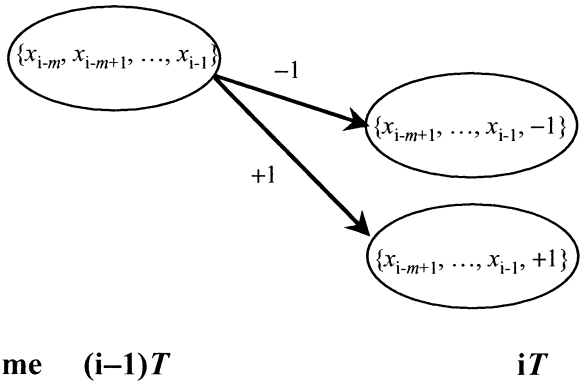


FIGURE 13 Diagram showing state transitions.

Since there are m interfering symbols and each symbol can take on two possible values, the intersymbol interference takes on 2^m possible values. By keeping track of the previous m symbols, it is possible to determine the amount of interference occurring in the current bit interval. The different combinations of the m interfering symbols, $\{x_{i-m}, x_{i-m+1}, \ldots, x_{i-1}\}$, are referred to as the channel states.

When the current input symbol x_i is transmitted, the channel transitions from state $\{x_{i-m}, x_{i-m+1}, \ldots, x_{i-1}\}$ to state $\{x_{i-m+1}, x_{i-m+2}, \ldots, x_i\}$. As shown in Fig. 13, there are two branches leaving each state corresponding to the two possible values of x_i.

Since the number of channel states increases exponentially with the degree of the partial response polynomial, m, the complexity of a detector that tracks the intersymbol interference over all possible channel states grows in a same manner.

C. Maximum-Likelihood Detection in Recording Systems

The channel bits associated with a partial response signal are detected using a maximum-likelihood (ML) detector. A ML detector finds the most likely input sequence given the entire sequence of samples, $\underline{r}$. A ML detector delays the decision-making process until all the samples, past, present and future, are considered.

In recording systems, ML detectors are generally designed assuming that the noise samples are independent uncorrelated random variables. While this keeps detector complexity manageable, it is well known that this simplification ignores the complex nature of the noise in recording systems.

Media noise is not only correlated it is signal dependent. In addition, the linear filtering in the post-equalizer

correlates the electronic noise that was previously uncorrelated. The electronic noise correlation is minimized when the partial response polynomial is close to the natural shape of the readback signal. Since incorporating the true noise characteristics into the detector is difficult, the detector assumes that the noise is independent and uncorrelated and suffers a SNR penalty due to noise correlation.

When the noise is Gaussian with zero mean and variance σ^2, each sample r_i is a Gaussian random variable, with mean:

$$y_i = \sum_{j=0}^{m} h_j x_{i-j}$$

and variance σ^2. For a given input sequence $\underline{x}$, the probability of a particular vector of n samples occurring is the product of n Gaussian probability density functions:

$$Pr\{\underline{r} \mid \underline{x}\} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(r_i - y_i)^2\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(r_i - y_i)^2\right)$$

The most likely input sequence is the one that maximizes the probability of $\underline{x}$ given $\underline{r}$, $Pr\{\underline{x} \mid \underline{r}\}$.

Assuming each input sequence has equal prior probability, the input sequence that maximizes $Pr\{\underline{x} \mid \underline{r}\}$ also maximizes $Pr\{\underline{r} \mid \underline{x}\}$, since the two probabilities are proportional to each other. Furthermore, maximizing $Pr\{\underline{r} \mid \underline{x}\}$ is equivalent to minimizing $-\log_e Pr\{\underline{r} \mid \underline{x}\}$. Therefore, a ML detector finds the input sequence $\underline{\hat{x}}$ that minimizes the squared Euclidean distance between the noisy samples and the ideal samples associated with an input sequence $\underline{x}$:

$$\sum_{i=1}^{n}(r_i - y_i)^2$$

1. Probability of an Error Event for a ML Detector

An error event occurs when the squared Euclidean distance for the correct input sequence is larger than the squared Euclidean distance for the most likely input sequence:

$$\sum_{i=1}^{n}(r_i - y_i)^2 > \sum_{i=1}^{n}(r_i - \hat{y}_i)^2$$

Let d^2 represent the squared Euclidean distance between the ideal values on the two paths:

$$d^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

then the probability of an error event occurring is given by:

$$Pr\left\{\sum_{i=1}^{n} n_i(\hat{y}_i - y_i) > \frac{d^2}{2}\right\} = Q\left(\frac{d}{2\sigma}\right)$$

where the Q function is defined as:

$$Q(z) = \int_z^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

Since the Q function falls off exponentially, the error events with the minimum distance, d_{min}, happen much more frequently than error events with larger distance. Therefore, the error event rate for a ML detector is dominated by the probability of a minimum distance error event:

$$Q\left(\frac{d_{min}}{2\sigma}\right)$$

2. Viterbi Algorithm

A maximum-likelihood detector searches over all possible input sequences using an efficient recursive algorithm known as the Viterbi algorithm. A trellis is used to represent all possible input sequences (see Fig. 14). Each path through the trellis represents a different binary input sequence. A trellis that extends for n sampling times contains unique paths for all 2^n possible input sequences.

Each branch in the trellis has associated with it an input symbol x_i, an ideal value y_i, and a branch metric equal to $(r_i - y_i)^2$. The sum of the branch metrics along a path, referred to as the path metric, represents the squared Euclidean distance between the received samples and the ideal samples associated with that path.

The branch metrics can be simplified by noting that r_i^2 is common to all the branch metrics. Consequently, minimizing the squared Euclidean distance is equivalent to minimizing:

$$\sum_{i=1}^{n}(r_i - y_i)^2 - r_i^2 = \sum_{i=1}^{n} 2r_i y_i + y_i^2$$

The Viterbi algorithm systematically eliminates those paths that cannot be part of the most likely path because they diverge and remerge with another path that has a smaller path metric. Since all other paths have been eliminated, a ML detector only needs to keep track of the minimum metric path leading to each state at the current sampling time. The metric of the minimum metric path is commonly referred to as the state metric. When the current sample is received, path metrics for the two paths leaving each state at the previous sampling time are calculated by adding branch metrics to state metrics. Then, the two path metrics entering each state at the current sampling time are compared and the path with the minimum metric is selected as the survivor path.

After all n samples have been received, the path with the overall minimum path metric is selected as the most likely path and the input sequence associated with that path is the maximum-likelihood sequence estimate $\underline{\hat{x}}$.

3. Partial Response Class 4

The first generation of partial response maximum likelihood (PRML) detectors implemented in hard disk drives used the partial response polynomial $h(D) = 1 - D^2$, which is referred to as partial response class 4 (PR4). For the linear densities used at that time, the PR4 pulse response was a good match to the natural shape of the readback. Since the PR4 polynomial has degree $m = 2$, there are four states corresponding to the four possible combinations of $\{x_{i-2}, x_{i-1}\}$. The PR4 signal at sampling time iT is given by:

$$r_i = y_i + n_i = (x_i - x_{i-2}) + n_i$$

where y_i is the ideal (noiseless) sample value and n_i is the noise. There are three possible ideal values: -2, 0, and $+2$.

The four-state trellis for the PR4 channel is shown in Fig. 14, with branches labeled x_i/y_i. The dashed lines in the trellis illustrate a minimum distance error event of length three that diverges from state 3 at time k and remerges at state 2 at time k + 3.

The minimum distance error events on the PR4 trellis have squared Euclidean distance $d^2_{\min} = 8$. Therefore, the probability of an error-event is given by:

$$Q\left(\frac{\sqrt{2}}{\sigma}\right)$$

The minimum distance error events for the PR4 channel are shown in Table I. As indicated by the dots in the last row in the table, the minimum distance error events can have infinite length since they only gain distance when they diverge and remerge. As a consequence, a minimum distance error event can cause an infinite number of channel bit errors in the estimate $\underline{\hat{x}}$.

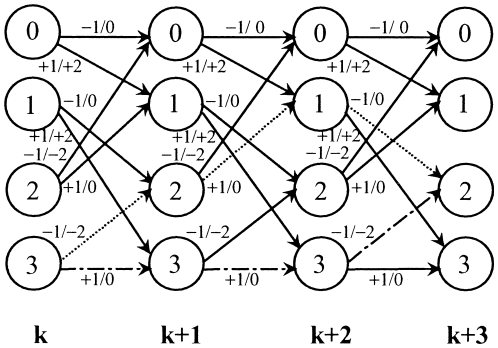


FIGURE 14 The four-state trellis for the PR4 channel.

TABLE I Minimum Distance Error Events for the Precoded PR4 Channel

$\underline{v} - \underline{\hat{v}}$	$\underline{w} - \underline{\hat{w}} = \frac{\underline{x} - \underline{\hat{x}}}{2}$	$\underline{y} - \underline{\hat{y}}$
±10±1	±1	±(20−2)
±1000±1	±(101)	±(2000−2)
±100000±1	±(10101)	±(200000−2)
±10000000±1	±(1010101)	±(20000000−2)
...	...	...

A precoder is often used to limit the number of bit errors per minimum distance error event. For the PR4 channel, a $1/(1 \oplus D^2)$ precoder limits the number of bit errors to two, where $\oplus$ denotes addition modulo 2. Let $\underline{v} = \{v_1, \ldots, v_n\}$ denote the binary input to the precoder, $v_i \in \{0, 1\}$. Then, the output of the precoder is $\underline{w}$, where:

$$w_i = v_i \oplus w_{i-2}$$

The sequence $\underline{w}$ is a binary version of the channel input sequence $\underline{x}$. A block diagram of the precoded PR4 channel is shown in Fig. 15.

The PRML detector typically outputs a binary estimate of the channel input, $\underline{\hat{w}}$. When the channel is precoded, a postcoder that performs the inverse of the precoder is used to obtain an estimate of the precoder input, $\underline{\hat{v}}$. For the precoded PR4 channel, the postcoder is $(1 \oplus D^2)$ such that:

$$\hat{v}_i = \hat{w}_i \oplus \hat{w}_{i-2}$$

4. PR4 ML Detector

A convenient property of the PR4 channel is that the even-numbered input bits interfere with each other but are independent of the odd-numbered input bits. Therefore, the PR4 channel can be represented as two interleaved $h(D') = (1 - D')$ channels, where D' is a $2T$ delay operator. As a consequence, the PR4 maximum-likelihood detector can be implemented as two interleaved $(1 - D')$ detectors. Since each of the interleaved detectors is clocked at half-rate, this implementation allows the detector to run twice as fast as a standard maximum-likelihood detector.

The $(1 - D)$ channel is a very simple two-state channel, where the states represent the two possible values of x_{i-1}. The two-state trellis diagram for a $(1 - D)$ channel is shown in Fig. 16, with branches labeled with $x_i/(y_i = x_i - x_{i-1})$.

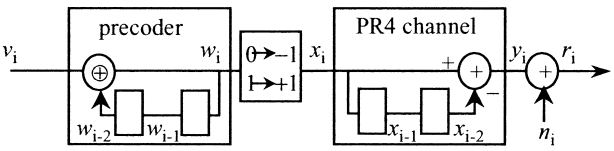


FIGURE 15 Block diagram of a precoded PR4 channel.

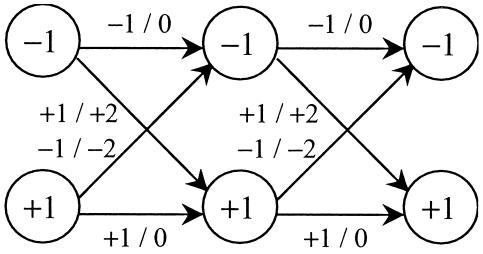


FIGURE 16 Trellis diagram for a $(1 - D)$ channel.

The ideal values take on three possible ideal values -2, 0, and $+2$. However, not all combinations of the three ideal values are allowed. In fact, the $+2$'s and -2's must alternate in the three-level ideal sequence. Given this constraint, the ML detector searches the received sequence for the most likely positions of the $+2$'s and -2's. After receiving sample r_i, a ML detector for a $(1 - D)$ channel calculates four branch metrics, using the simplified branch metric $(-2r_i y_i + y_i^2)/4$. Since two of the branch metrics are zero, the detector only needs to calculate $r_i + 1$ and $-r_i + 1$.

Next, the state metrics are updated using a process called an add–compare–select. The first step in the process involves calculating path metrics. The second step involves comparing the two path metrics entering each state. Finally, the third step is to select the survivor paths. The path metric associated with the transition from state x_{i-1} to state x_i, $P_i(x_{i-1}, x_i)$ is obtained by adding the branch metric to the state metric, $M_{i-1}(x_{i-1})$. The four path metrics are given by:

$$P_i(-1, -1) = M_{i-1}(-1)$$
$$P_i(-1, +1) = M_{i-1}(-1) - r_i + 1$$
$$P_i(+1, -1) = M_{i-1}(+1) + r_i + 1$$
$$P_i(+1, +1) = M_{i-1}(+1)$$

After comparing the incoming path metrics, the new state metrics are given by:

$$\begin{aligned} M(-1) &= \min\{P_i(-1, -1), P_i(+1, -1)\} \\ &= P_i(s_i(-1), -1) \\ M(+1) &= \min\{P_i(-1, +1), P_i(+1, +1)\} \\ &= P_i(s_i(+1), -1) \end{aligned}$$

where the output of the comparitors is expressed as:

$$s_i(-1) = \text{sign}\{P_i(-1, -1) - P_i(+1, -1)\}$$
$$s_i(+1) = \text{sign}\{P_i(-1, +1) - P_i(+1, +1)\}$$

Since new survivor paths have been selected, the survivor path memory that contains a record of the input bits along each of the survivor paths must be updated. The updating process involves using the output of the comparitors to select and extend one of the old survivor paths. Let $S_i(x_i)$ denote the survivor path for state x_i, then

$$s_i(-1) = \{S_{i-1}(s_i(-1)), -1\}$$
$$s_i(+1) = \{S_{i-1}(s_i(+1)), +1\}$$

For the $(1 - D)$ channel, there is an alternative implementation of the ML detector that uses the dynamic threshold algorithm rather than the Viterbi algorithm. This equivalent approach provides an intuitive understanding of how the ML detector searches the received sequence for the most likely positions of the alternating $+2$'s and -2's. The dynamic threshold algorithm is derived by rearranging the terms in the compare, so that the sample r_i is on one side and a threshold is on the other side. Since there is one compare for each of the two states, there are two thresholds. The upper threshold is $\Delta_{i-1} + 1$ and the lower threshold is $\Delta_{i-1} - 1$, expressed in terms of the difference between the state metrics:

$$\Delta_{i-1} = M_{i-1}(-1) - M_{i-1}(+1)$$

The two dynamic thresholds move up and down in response to the received sequence, while keeping a constant distance of two apart from each other, as shown in Fig. 17.

The detector starts out with an upper threshold at $+1$ and a lower threshold at -1, since $\Delta_0 = 0$. Samples that fall in between the two thresholds are immediately determined to have an ideal value of 0. A sample that is greater than the upper threshold is considered the most likely $+2$ seen since the last -2. In response, the thresholds are moved up so that the upper threshold is equal to the sample. Similarly, a sample that is less than the lower threshold is considered the most likely -2 seen since the last $+2$. In response, the thresholds are moved down so that the lower threshold is equal to the sample. A final decision on the best $+2$ is made when the first possible -2 is encountered, and vice versa. After receiving sample r_i, the difference metric is updated directly, without calculating the individual state metrics:

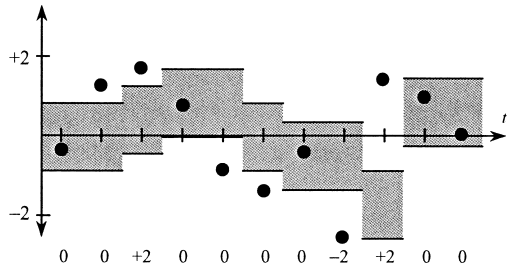


FIGURE 17 Dynamic threshold algorithm for a $(1 - D)$ channel.

$$\Delta_i = \begin{cases} r_i + 1 & \text{if } r_i < \Delta_{i-1} - 1 \\ \Delta_{i-1} & \text{if } \Delta_{i-1} - 1 < r_i < \Delta_{i-1} + 1 \\ r_i + 1 & \text{if } \Delta_{i-1} + 1 < r_i \end{cases}$$

The survivor paths are updated in exactly the same way as in the Viterbi algorithm:

$$s_i(-1) = \text{sign}\{\Delta_{i-1} - 1 - r_i\}$$
$$s_i(+1) = \text{sign}\{\Delta_{i-1} + 1 - r_i\}.$$

5. Extended Partial Response Class 4

The second generation of partial response channels implemented in disk drives used partial response polynomial $h(D) = 1 + D - D^2 - D^3$, which is referred to as extended partial response class 4 (EPR4). The minimum distance error events on the EPR4 trellis have squared Euclidean distance $d^2_{\min} = 16$. Due to the doubling of the minimum squared Euclidean distance, the EPR4 channel provides 3 dB more SNR than the PR4 channel. In addition, the EPR4 channel is a better match to the shape of the readback signal than is the PR4 channel at higher linear densities. Since there is less post-equalization, there is less of a SNR penalty due to noise correlation ignored in the detector.

The two types of minimum distance error-events for the EPR4 channel are shown in Table II. The rows of dots in the table indicate that there are infinitely long minimum distance error events that can cause an infinite number of channel bit errors in the estimate $\underline{\hat{x}}$. For the EPR4 channel, a $1/(1 \oplus D^2)$ precoder limits the number of bit errors associated with a minimum distance error event to two or four.

Since there are three interfering symbols, the EPR4 channel has eight states corresponding to the eight possible combinations of $\{x_{i-3}, x_{i-2}, x_{i-1}\}$. The ML detector for the EPR4 channel has twice the complexity of a PR4 detector and runs half as fast, since the EPR4 detector must run at full rate.

TABLE II The minimum distance error-events for the precoded EPR4 channel

$\underline{v} - \underline{\hat{v}}$	$\underline{w} - \underline{\hat{w}} = \frac{\underline{x} - \underline{\hat{x}}}{2}$	$\underline{y} - \underline{\hat{y}}$
$\pm 1\,0 \pm 1$	± 1	$\pm(2\,2\,-2\,-2)$
$\pm 1\,0\,0\,0 \pm 1$	$\pm(1\,0\,1)$	$\pm(2\,2\,0\,0\,-2\,-2)$
$\pm 1\,0\,0\,0\,0\,0 \pm 1$	$\pm(1\,0\,1\,0\,1)$	$\pm(2\,2\,0\,0\,0\,0\,-2\,-2)$
...	...	...
$\pm 1 \pm 1\,0 \pm 1 \pm 1$	$\pm(1\,-1\,1)$	$\pm 1(2\,0\,-2\,2\,0\,-2)$
$\pm 1 \pm 1\,0\,0 \pm 1 \pm 1$	$\pm(1\,-1\,1\,-1)$	$\pm 1(2\,0\,-2\,0\,-2\,0\,2)$
$\pm 1 \pm 1\,0\,0\,0 \pm 1 \pm 1$	$\pm(1\,-1\,1\,-1\,1)$	$\pm 1(2\,0\,-2\,0\,0\,2\,0\,-2)$
...	...	...

D. Modulation Coding for PRML

In partial response channels with infinitely long minimum distance error events, a modulation code is used to limit the length of the minimum distance error events so that the decision latency in the PRML detector is finite. Since the performance of the detector is dominated by minimum distance error events, the survivor path memory can be implemented with a fixed decision delay that is greater than or equal to the length of the longest minimum distance error event.

The minimum distance error events for the precoded PR4 channel correspond to having a long run of 0's in one interleave at the input to the precoder. The minimum distance error events for the EPR4 channel correspond to the precoder input containing a long run of 0's in one interleave or having long runs of 0's in both interleaves simultaneously, resulting in a long run of 0's in the global sequence.

A (G, I) code is the type of RLL modulation code used with PRML. A (G, I) code has at most G 0's in between two 1's in the global sequence and has at most I 0's in between two 1's in the even and odd interleaved sequences. Note that an interleave constraint of I implicitly provides a global constraint of $2I$.

When applied to a precoded PR4 channel, the longest error event has length $2I + 3$. When applied to a precoded EPR4 channel, the longest error-event has length $2I + 4$, if $G < 2I$, and length $2I + 5$, if $G = 2I$. In addition to reducing the decision latency, the global constraint in the modulation code ensures that transitions occur frequently which guarantees that there will be enough timing information to track changes in phase and frequency.

E. Error-Correction Coding

The error-correction coding is designed to handle the random errors associated with electronic and media noise as well as the burst errors associated with thermal asperities, drop-outs, or media defects.

In most cases, the ECC is a systematic code, which means that redundant parity bits are calculated and appended to the end of each codeword. The rate of an ECC code is the number of user bits divided by the total number of bits in a codeword.

Most digital magnetic recording systems use (n, k) Reed–Solomon codes with 8-bit symbols, where n is the number of symbols in a codeword, k is the number of user symbols, and $(n - k)$ is the number of redundant parity symbols. The code rate is k/n. The error-correction capability for this code is $t = (n - k)/2$. Any combination of t or fewer symbol errors in a codeword is corrected during decoding.

Since the host computer organizes the user data into 8-bit bytes, the Reed–Solomon code usually operates on 8-bit symbols. However, this is not a requirement. Reed–Solomon codes can operate on symbols composed of any number of bits. For a given code rate, a larger symbol size provides more error correction.

One of the drawbacks to 8-bit symbols is that the maximum codeword length is $n = 2^8 - 1 = 255$. Since the host computer sends user data in 512-byte sectors, one sector is split into three codewords. Since each codeword can correct t byte errors, the three codewords are interleaved so that a solid burst of $3t$ byte errors can be corrected. If longer burst error correction is required, the number of interleaved ECC codewords can be increased. On the downside, interleaving codewords is not very efficient. The number of parity symbols increases but the probability of a sector failing stays the same, since $t + 1$ or more byte errors in any of the codewords cause a sector failure.

When a sector fails, the ECC usually detects that there were too many errors. In this case, the sector is reread as many times as it takes to decode the sector. Since rereading a sector involves waiting for the disk to complete a revolution, the probability of a reread must be small so that the average seek time is not impacted. The worst type of failure occurs when there are so many errors that the ECC decodes to the wrong codeword without detecting a failure. To reduce the probability of miscorrection, an additional cross-check code is typically used to detect this type of failure. The probability of miscorrection must be so small that it would not happen over the lifetime of the disk drive.

Considering that the heads and media in a disk drive are designed so that the error event rate at the output of the detector is on the order of $1e-5$, the error-correction coding does an admirable job of maintaining a mean time between failures (MTBF) of approximately 400,000 hours.

ACKNOWLEDGMENT

The author wishes to thank James Fitzpatrick for reviewing the manuscript.

SEE ALSO THE FOLLOWING ARTICLES

DESIGNS AND ERROR-CORRECTING CODES • DIGITAL ELECTRONIC CIRCUITS • DIGITAL SPEECH PROCESSING • MAGNETIC MATERIALS • MUSICAL ACOUSTICS • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, DIGITAL

BIBLIOGRAPHY

Adler, R., Coppersmith, D., and Hassner, M. (1983).“Algorithms for sliding block codes,” *IEEE Trans. on Information Theory* **29**(1), 5–22.

Dolby, R. M. (1967) “An audio noise reduction system,” *J. AES* **15**(4), 383–388.

Ferguson, M. J. (1972) “Optimal reception for binary partial response channels,” *Bell Systems Tech. J*., **Feb**. 493–505.

Forney, G. D. (1972). “Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference,” *IEEE Trans. on Infomation Theory* **18**(3), 363–378.

Grochowski, E., and Thompson, D. A. (1994). “Outlook for maintaining areal density growth in magnetic recording,” *IEEE Trans. on Magnetics* **30**(6), 3797–3800.

Kabal, P., and Pasupathy, S. (1975). “Partial-response signaling,” *IEEE Trans. on Commun*. **23**(9), 921–933.

Lin, S., and Costello, Jr., D. J. (1983). “Error Control Coding: Fundamental and Applications,” Prentice-Hall International, Englewood Cliffs, NJ.

Mee, C. D., and Daniel, E. D. (1990). “Magnetic Recording Handbook,” McGraw-Hill, New York.

Mintzer, F., and Howell, T. (1982). “Experiments on the Use of FDDT for AC-Bias Disk Recording,” IBM-RC 9429 (#41644). June 16.

Siegel, P. H., Wolf, J. K., and Marcus, B. (1992). “Finite-state modulation codes for data storage,” *IEEE J. Selected Areas Commun*. **10**(1), 5–37.

Thapar, H. K., and Patel, A. M. (1987) “A class of partial response systems for increasing storage density in magnetic recording,” *IEEE Trans. on Magnetics* **23**(5), 3666–3668.

Viterbi, A. J. (1967). “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.” *IEEE Trans. on Information Theory* **13,** 260–269.

Wallace, R. L. (1951). “The reproduction of magnetically recorded signals,” *Bell Systems Tech. J*. **30,** 1145.

Wood, R., and Peterson, D. (1986). “Viterbi detection of class IV partial response on a magnetic recording channel,” *IEEE Trans*. on Commun. **34**(5), 454–461.



Microwave Communications

U. S. Berger
Bell Telephone Laboratories (retired)

I. Introduction
II. Early Work
III. Early Needs and Applications
IV. Microwave System Design
V. Waveguide
VI. Antennas
VII. Components
VIII. Reliability
IX. Radio Terminals
X. System Testing
XI. Single-Sideband AM Radio
XII. Satellite Communication
XIII. Digital Microwave Radio
XIV. Conclusion

GLOSSARY

Bandwidth Frequency occupancy needed for the satisfactory transmission of useful information. Very often the frequency spectrum is greater than that theoretically needed.

Baseband Band of frequencies, usually the lowest frequencies in the microwave system, where basic information is assembled; this spectrum is generally that provided to the microwave system to be delivered to a distant point in the same format and information content.

Carrier Used with two widely different, yet unambiguous meanings: (1) in business, organizational entity to render communications to customers or users; (2) electrical wave which by itself does not carry information but is generally modulated by another wave which does contain desirable information.

Compandor In telephony, use of a compandor (a device that reduces the amplitude level of strong talk spurts) at the head end of a system permits the average level of the voice channel to belevel; raised without exceeding the system design this automatically raises the level of the weak talk spurts in relation to the stronger speech signals. At the receiving end, the inverse is done (expandor) to restore the original relationship of the strong spurts to the weak ones. Overall effect is a very substantial improvement in the signal-to-noise ratio on voice circuits.

Dielectric resonators (YIG) Magnetic devices that have certain desirable characteristics, primarily magnetic resonances, which can be used for frequency selective elements in microwave networks (filters, oscillators, equalizers, etc.).

Digital transmission Signals sent as a series of pulses of a prescribed amplitude and received and detected in predetermined time slots. Digital signals for microwave transmission are usually encoded and processed in such a way that the spectral bandwidth is minimized.

Downconverter (or mixer) In a superheterodyne receiver, the circuit that accepts the SHF signal and the channel carrier to produce a difference frequency known as the intermediate frequency.

Drop and add In a long, multichannel microwave

communication facility not all of the information is conveyed from one end of the system to the other: signals may be dropped to customers along the way (e.g., network television). Similarly, the channel may accept new (added) signals at intermediate points. Access to the system baseband is necessary for this function.

Hot standby Second repeater tuned to the same frequency as another and available for substitution into the system for the working equipment that has failed. Substitution is usually made automatically after a failure has been detected.

Intermediate frequency (IF) Frequency, in the vicinity of 70 MHz (which is a common frequency for all channels) at which considerable amplification takes place, interconnections are made, automatic gain adjustment is provided, and channels are disabled upon command (squelch). Use of the relatively low frequency simplifies the design of these functions compared to performing them at superhigh frequency (SHF). IF is derived by demodulating the SHF signal with a channel carrier.

Intermodulation Heterodyning or undesired mixing of two signals in a nonlinear device to produce an undesired product.

Intersymbol interference In digital systems, distortion can cause a part of one pulse to overlap other pulses so that it may be difficult or impossible to identify the presence or absence of a specific pulse of interest.

Isotropic antenna Theoretical antenna of infinitesimal size in which it is assumed that all of the energy is radiated (point source). It has no visual shape or form and has no practical use except as a basis for reference for other antennas of finite dimensions.

Klystron Electron tube that can operate in the SHF bands. Some klystrons are designed as oscillators with a single cavity (reflex klystrons). Multicavity klystrons can amplify or oscillate, depending upon the configuration of the external components used with the device.

Modulation and demodulation Process whereby useful information is translated from one portion of the frequency spectrum to another. Demodulation is the inverse of modulation. Modulation is sometimes used interchangeably with detection. Modulation is frequently used to connote the mixing of two frequencies to derive a third frequency.

Negative resistance Only positive resistances exist as discrete passive components that can be purchased to specific values from a vendor. Positive resistances obey Ohm's law, where $R = E/I$. This means that the component is linear; higher voltage across it will result in higher current through the device. In a negative resistance, $R = -E/kI$, which states that over some range of values an increase in voltage will result in a reduction in current. This relationship will maintain only over a limited range of the E–I characteristic.

Noise figure Measure of the excess noise in a device over the theoretical or ideal conditions. Very useful measure of the quality of an active component in a radio system. The lower the noise figure, the better the performance and the less contribution the component will make to system noise.

Polarization Relation of the electromagnetic wave vectors to some reference. Polarization may be vertical (wherein the electric vectors are in a vertical plane), horizontal, circular (right or left handed), and so on. Coupling between H and V polarizations can be substantially minimized with careful design, and this can be helpful in reducing adjacent channel couplings in multichannel systems. 30- to 35-dB cross-polarization discrimination can be obtained in some cases.

Repeater Radio station capable of receiving a weak signal on a specific frequency and substantially increasing the power for transmission on another frequency within the system spectrum.

Superheterodyne Type of radio equipment in which an incoming frequency is converted to another frequency for purposes of amplification and for providing selectivity against unwanted frequencies. A carrier frequency differing from the incoming signal by a specific amount, the IF frequency, and coupled into a nonlinear device produces the modulation function.

Superhigh frequency (SHF) Wavelength shorter than 10 cm; frequencies above 3 GHz.

Waveguide Usually a hollow metal structure intended to guide an electromagnetic wave along its length. Internal dimensions are related to the efficacy of transmission at specific frequencies. Cutoff frequency is the frequency below which a particular waveguide cannot satisfactorily transmit the wave.

Waveguide modes Depending on the shape and size of the waveguide with respect to the length of the wave traversing the guide, one or more patterns of wave propagation may exist within the guide. Each of these modes will have a specific topology, different velocity, and different energy distribution along and across the guide cross section. Both desirable and undesirable modes can exist under specific conditions.

THE TRANSMISSION of television, telephone, and data services via superhigh radio frequencies is known as microwave communications and has been available since the late 1940s. Terrestrial microwave systems consist of a series of radio stations, generally located within line of sight of the adjacent stations, where the signals may be retransmitted repeatly to achieve a system of substantial length.

This has frequently been referred to as a microwave radio relay system or a microwave repeater system. Usually, these facilities are capable of carrying a number of TV programs or thousands of telephone or data circuits simultaneously. Microwave communication systems have become highly reliable, efficient, and economical in handling communications of all types, and a vast network of circuits has been provided since the late 1940s. Satellite microwave systems have supplemented the terrestrial facilities in providing worldwide communications.

I. INTRODUCTION

Communication via microwave radio is not a recent development for theoretical studies of radio transmission in the superhigh frequency (SHF) band (3 GHz and above) and was carried out in the late 1930s. Throughout the world, scientists and engineers were pushing the frontiers of all radio transmission higher and higher in frequency. Well before World War II radio communication was being vigorously pursued in the frequency region from 50 to 500 MHz for frequency modulation (FM) broadcasting, point-to-point transmission, television broadcasting, and vehicular communication (including aircraft). Military aircraft radio communication in this portion of the spectrum was highly successful in the 1940s and aided in the defeat of the Japanese in the Pacific because they were unaware of our substantial communication capabilities at these higher frequencies. Television broadcasting which requires substantial bandwidths in the radio spectrum, was demonstrated to the public at the New York World's Fair in 1939. All of these were part of the evolution of the field of communication that started in the 1930s and has continued until the present with communications people today working at lightwave frequencies with glass fibers as the transmission medium.

II. EARLY WORK

Researchers in many countries in the 1930s were exploring radio at even higher frequencies, and many basic concepts pertaining to SHF were well understood before World War II. The lack of adequate components and test equipment slowed the progress of these studies, and it was often necessary to pause and invent or make the parts that were needed.

At frequencies above $\sim$3 GHz, some of the components and subunits within a microwave station are interconnected with waveguide, but in the early days the waveguide itself was equally unknown and had to be thoroughly investigated. One of the research pioneers in this field was George Southworth of Bell Laboratories. He and his research team evaluated the loss and other characteristics of a radio wave as it was propagated through a smooth metal pipe. To obtain meaningful results without a prohibitively long section of pipe, a very short pulse of radio energy was injected into the pipe and allowed to traverse back and forth many times until the loss in the walls of the waveguide attenuated the signal below a detectable level. Counting the detected pulses, measuring the time required for successive pulses to pass the detection point, and observing the amplitude of the pulse each time it appeared at the detector, the characteristics of the medium could be determined. Other equally significant studies of microwave signals were conducted by many people during this period.

As a result of the work done before 1940, it was possible to develop a four-channel microwave system for the Armed Services operating in the 4.3- to 4.8-GHz band using a reflex klystron tube with an output power of 0.1 W. This system was capable of carrying initially 8 voice circuits using pulse-position modulation with time multiplexing of the 8 circuits in the pulse stream (before the end of the war the capacity had been increased to 16 circuits). This early system was used in the field in Europe from 1943 through 1945 and provided communication over impassable terrain. With the pulse signal format, this system was the forerunner of what today is one of the newest and most important forms of communication, namely digital transmission. During and after the war, tests were made in the field to study propagation characteristics as they might be experienced in actual communication systems. A 40-mile link in the New York–New Jersey area was established, and propagation tests were made at a number of frequencies over a frequency range of 1 to 24 GHz. These results were extremely important in determining how a relay system might perform in a field environment and laid the basis for planning long-repeatered systems.

III. EARLY NEEDS AND APPLICATIONS

At the close of World War II, there was a backlog of needs for all forms of consumer goods, one of which was television. The above-mentioned demonstrations at the World of Tomorrow Exhibition in New York and the promise by the television people that their offering was about ready for widespread release to the public raised the question as to how the program material could be distributed to local broadcasters. While television could be transmitted via coaxial cable, the limited amount of cable available in the ground, the high cost of laying additional cable, and the length of time required to put new facilities in place throughout the country made microwave radio a very attractive alternative. The high cost of a cable system was

due to (1) the need for rights-of-way for the entire length of the cable trenches; (2) the high construction costs of routes through rocky terrain, particularly in the East, the Northeast, and the Rocky Mountain regions; and (3) the cost of the cable itself. Radio sites, with average spacings of about 25 miles on high hills and often in remote areas, were much more readily obtainable and at lower initial cost.

In 1944, plans were made for the development of an experimental microwave radio relay system between New York and Boston that could provide television service between the two cities, but more importantly, could also answer a myriad of very important questions on economics, expected quality of transmission, reliability, maintenance requirements, public acceptance, entrance into cities, and relations with the entertainment industry. Earlier studies had already addressed the question of what frequency band should be used for this type of service, and it was concluded that somewhere between 3 and 10 GHz would be acceptable from a propagation standpoint. From the component point of view, the lower end of this range would be preferable for the first system. Below 3 GHz, components tend to become large and cumbersome, whereas frequencies above 9 or 10 GHz are subject to severe attenuation problems in areas of high rainfall such as in the Gulf Coast region. Negotiations with the FCC led initially to a common carrier frequency allocation of 3.9 to 4.2 GHz, but this was later expanded to 3.7 to 4.2 GHz. Other bands at 6 and 11 GHz were explored but developed later.

The four two-way channels with seven intermediate repeater stations and the terminal stations in New York and Boston represented a substantial amount of microwave hardware that had to be manufactured and installed. Including all of the ancillary items such as antennas, buildings, and power and site preparation, the aggregate added up to a rather expensive experiment. The equipment units in the New York–Boston system included klystron oscillators and klystron amplifiers to boost the power to approximately 0.5 W. Narrowband FM was selected for the modulation plan to avoid the nonlinear amplitude distortion effects of the klystron amplifiers, which would introduce intolerable distortions in the television picture and in the telephony circuits that were to be studied during the early testing. The New York–Boston SHF radio link was completed in November of 1947 and was turned up for use by the television industry for their experiments. In May of 1948, commercial service between the two cities was inaugurated.

While the New York–Boston system was being constructed, plans were being made to build a much longer 4-GHz facility to serve a larger portion of the country. New York to Chicago was the first phase to be built through New York State, along the Great Lakes in Ohio and Indiana, and into Chicago. This flat farm country was easy terrain on which to build a radio system. The route was expedited for the 1950 World Series, and this was a treat for those television viewers along the way who previously had only local programs or film presentations to watch. It would not be correct to say that there were no troubles in the New York–Chicago route: equipment failures, particularly failures of electron tubes, were numerous and a major source of trouble, but this may not be surprising when the number of components operating in tandem is taken into account. There were about 20 electron tubes in each repeater and about 35 repeaters in the route for a total of 700 tubes, many in unattended locations. Component life then was not good, but even with a mean life of 5000 hr for the tubes, a failure could be expected every 7 hr. Not knowing which tube would fail next, the total downtime of the system was appreciable and the maintenance effort was high. These were the growing pains of a new venture.

Despite these kinds of reliability problems and some performance difficulties, the second phase of the transcontinental route was being planned before the troubles in the first section were all solved. The Chicago–San Francisco route was more difficult to build through the mountainous areas, but sites could be readily found in the farm belt of the Midwest and the mountaintops of the Rockies. Service dates were targeted for the San Francisco peace talks with Japan in September of 1951, and with about 65 repeater sites to be built, installed, and tested, the task was enormous, but again the deadline was met. Reliability continued to be a problem, and like any first experience, many of these difficulties were overcome by careful study, analysis of the failure information, redesign of parts that were causing trouble, and improved manufacturing techniques of many components. In Section VIII the subject of reliability is addressed in greater detail. While the backbone of a network was being formed, side branches were also being planned and constructed in many parts of the country, and this has continued over the years until the present; virtually every locality is being served by microwave radio.

IV. MICROWAVE SYSTEM DESIGN

Some of the considerations that go into the design of a microwave system include (1) the service needs; (2) the initial size of the route in terms of length and channel cross section; (3) the potential for growth; (4) the number of drop-and-add points where service is to be supplied; (5) the availability of frequencies in the vicinity of the proposed route; (6) the availability of suitable sites for the repeater stations; (7) the noise and reliability objectives; (8) the availability of power and access roads; and (9), not the least, cost.

There are two general types of systems in use: (1) the long-haul system, which may have only a modest number of drop-and-add points but may have many repeater stations and may go a long distance end to end; and (2) the short-haul system, which may consist of a relatively small number of repeaters with frequent drop-and-add points. Both of these systems use a superheterodyne type of receiver with an intermediate frequency of 65 or 70 MHz.

Three common carrier bands, 4, 6, and 11 GHz, have been used for communication purposes, and generally, the best overall use of these bands has been made by the communications industry. The early use of 4 GHz for long-haul service, such as between New York and San Francisco, more or less established a pattern that 4 GHz would be used for long haul. Similarly, the 11-GHz portion of the spectrum, which experiences problems in heavy-rainfall areas, would not be used in long-haul facilities. As might be expected, 6 GHz has been employed for both applications, and because short, light route sections of 6 GHz were built in a great many places in the United States, full development of long-haul 6-GHz routes was not always possible.

A. Short Haul

Short-haul systems are generally used where traffic loads are relatively light or the length of the route is not great, for television pickup from remote location for intrastate facilities for the common carriers, for control communications for gas and electric utilities, for telemetry control signals for dams, and for some government communications. Because the distances may not exceed a few hundred miles and the traffic load may be relatively light, the performance requirements have been set slightly below long-haul standards. The need for frequent access to and from the system has mandated detection and modulation capabilities in the repeater, and when these functions are provided, some distortion occurs. This type of repeater is known as a remodulating repeater, and the connection between the receiver and the transmitter within the repeater is made at baseband frequencies (0–5 MHz) through a level-adjusting network to assure that the proper deviation of the transmitter is obtained.

B. Long Haul

Long-haul systems, on the other hand, may have a number of repeaters that require no dropping or adding, primarily because they may be located in open country away from cities or towns where no baseband access to the system is required. It would be unwise to use remodulating equipment in these situations because the distortion would be too high and the deviation of the transmitters would be difficult to hold to the proper value. This leads to a second type of repeater, namely the heterodyne repeater, wherein the interconnection from the microwave receiver to the microwave transmitter is made at the 70-MHz IF rather than at baseband.

C. Growth

The service needs of a system, both initially and in the future, determine the number of radio channels that must be included in the design plans. Experience since the mid-1950s has shown that growth of microwave use in virtually all areas of the United States has occurred but not always at the same rate. In many cases, the original construction took care of only the initial requirements, and later, building and equipment additions were made to accommodate the growth needs. The 4-GHz band that was first developed, in many areas, has long been filled, and the facilities cannot be expanded further on those routes. The same can be said of the 6-GHz band; so, if these frequencies are to be reused for additional growth, new routes have to be constructed. This may be feasible in the open areas of the country, but eventually the routes must converge in the cities, where the service needs exist. The entrance problem has become an ever-increasingly serious one. In the period from the early 1950s until the present, microwave systems have grown into vast networks carrying all forms of communication services to all parts of the nation. Some telephone and data circuits traverse 2500 miles or more with little or no intermediate demodulation.

D. Performance

Noise and reliability objectives for the carriers are invariably set high because many customers are dependent on these facilities for important communications. A short "hit" or bat of noise may be of no consequence to a home viewer of a television program, for the eye is highly forgiving of a minor signal imperfection, or to a telephone user because a slight click in the telephone receiver is ignored, but the same "glitch" to an unforgiving computer may raise an order for 5 tons of steel to 5000 tons. (It was for this reason that data systems introduced error detection capabilities.) Noise and reliability enter into such crucial system design factors as transmitter output power, receiver noise figure, and repeater spacing.

Repeater spacing is determined from many considerations. Microwave transmission is almost always "line of sight" because this results in by far the best performance, and therefore high locations are almost always selected for the repeater station. Clearance over hills, trees, buildings, and other obstructions is sometimes difficult to obtain at a given point, and the designer may have to reexamine the terrain to get a radio path that is acceptable. If the distance

between repeaters is too great, the received signal may be too low and atmospheric fading may be excessive. However, route cost savings can result if repeaters are located far apart because less equipment will be needed. Conversely, short hops can improve both noise performance and reliability, but the cost will increase because more equipment will be required. Economics play an important role in system design. AT&T has a vast microwave network with an average hop length of about 26 miles. There are some hops as short as 1 mile because of terrain factors, for example, providing service into a valley from a repeater on a nearby mountain ridge. Repeaters over 50 miles apart to span some territory, such as an Indian reservation in the Southwest, are also found. A few long spans can be accommodated by the statistics of the performance of the system or by taking special steps, such as providing space-diversity antennas, to overcome serious fading phenomena.

E. System Capacity

In terrestrial FM microwave systems, only one television program can be carried on a single radio channel. There are a number of reasons for this: (1) the 20-MHz channel width at 4 GHz will only accommodate one television program; (2) the stacking and unstacking of a second television program above the first one in the baseband makes for very complex FM terminals; (3) interference between the two television programs on the same radio channel may put bars in the television pictures, and if the two pictures are not synchronized, the patterns will move; and (4) if the two signals are not coterminous, a television network management problem arises. Accordingly, television transmissions via microwave relay facilities are handled as a complete channel load and can be switched at IF if necessary.

Telephony, on the other hand, started with 480 circuits on a fully loaded 4-GHz radio channel in 1950, and as the technology developed, more and more circuits were added until, a quarter of a century later, the 4-GHz channels could each carry 1500–1800 4-kHz voice-frequency circuits (for either voice or data). The 6-GHz FM channels are about 30 MHz wide and can carry 1800–2400 voice-frequency circuits on each SHF channel.

When single-sideband (SSB) amplitude modulation (AM) at 6 GHz was turned up for service in early 1981, the voice circuit capacity was increased to 6000 circuits. It is quite apparent that if the basic components of a system, such as the antenna, waveguide, tower, basic and emergency power, station alarms, and land, can be used with a new system of greater capacity per radio channel, the individual voice circuits will be more economical and the development of a new and different kind of system can be justified.

F. Frequency Coordination and Channelization

Frequency coordination with other services and users in a specific area must be done to avoid serious interference problems and to provide for future growth for all users. The long-range plans of a carrier are taken into consideration when another service seeks to operate in the same band in the same area. Occasionally, the problem is adjudicated with the assistance of the FCC.

Very early in the history of microwave communication systems, extensive studies were made on frequency plans and channelization schemes and recommendations were made to the FCC for industrywide acceptance. Without standardization of a plan and the subsequent licensing of stations to use specific frequencies, utter chaos would exist. After the New York–Boston development was started, a frequency plan for the 3.7- to 4.2-GHz band was established using 24 20-MHz-wide channels with 12 transmitter frequencies interleaved with an equal number of receiving frequencies across the 500-MHz band. This scheme spread the high-level transmitting signals over the entire authorized 4-GHz band, and at that time, it was felt that this spreading would ease the receiver design problem. This channelization arrangement provided 12 usable 20-MHz frequencies in each of two directions of transmission. (Only a very small percentage of the licensed channels are used for one-way service. Studio transmitter links and remote television pickups are two examples.) An IF of precisely 70 MHz was chosen somewhat arbitrarily and, once selected, was retained for all 4-GHz equipment. This selection was satisfactory because, at that time, all repeater equipment used self-contained microwave-frequency-generating equipment.

When the 6-GHz common carrier band was developed starting about 1950, the plan placed eight 30-MHz transmitting channels in one-half of the 5.925- to 6.425-GHz spectrum and eight 30-MHz receiving channels in the other half, with a small guard band between. This was felt to be an advantage in systems using separate antennas for transmitting and receiving because the interferences would be less. It has turned out that these two plans are "fielder's choices" because, in later years, the same antenna has been used for both 4 and 6 GHz simultaneously. Hindsight indicates that only one scheme should exist, but unfortunately, once a decision is made and then is accepted as the standard, any deviation from it is impossible to introduce. The present arrangement is workable as attested to by the large numbers of 4- to 6-GHz joint routes, but without knowledge of the original arguments for one frequency plan versus the other, the justification for two plans to exist is virtually impossible to comprehend.

The first 6-GHz long-haul system used a common source of microwave frequency generation for all of the

repeaters in a relay station to minimize intermodulation tone problems. In essence, all channels were "synchronized" to a common source. To take full advantage of this asset, the IF was established at 74.1 MHz and the channel spacings were set at 29.6 MHz. This has become the industry standard on frequency assignments in this band and has also been adopted for the newest SSB AM radio systems.

V. WAVEGUIDE

Reference has already been made to the theoretical studies done on waveguide and its use in microwave systems. Waveguide is a component that, as the name implies, conducts or guides the microwave energy. Electromagnetic theory indicates that if the frequency is high enough the center conductor of a coaxial structure is no longer needed and the energy can be propagated on the inner walls of the medium. The cross section of the guide can be square, rectangular, circular, or elliptical. Each one of these configurations supports propagation in one or more modes, a mode being the way the electromagnetic energy is distributed within the pipe. Discrete modes or wave topologies are often used to obtain some specific result. The waveguide dimensions are somewhat critical for they have to be large enough to support the desired modes at the lowest frequency of interest, yet not be so large that undesired modes are also propagated. The waveguide size determines the cutoff frequency, a frequency below which satisfactory propagation cannot be achieved. Therefore, systems operating in different bands use different sizes of guides.

A. Rectangular Waveguide

Rectangular waveguide usually has a cross section with an aspect ratio of 1:2, the width being about twice the height. For operation in the 3.7- to 4.2-GHz band, the large inside dimension is 2.29 in. For the 10.7- to 11.7-GHz band, the wide dimension is 0.9 in. Waveguide is coded WR- (waveguide, rectangular) with numbers following to indicate the large dimension (i.e., WR-229). The numbers commonly used are the width dimension with the decimal point omitted.

The dominant mode in a rectangular guide is the TE_{01}, the transverse electric mode, where the electric vector is perpendicular to the direction of propagation in the guide. In this case, the electric wave vector is from one wall of the large dimension to the other.

The loss of waveguide is somewhat dependent on the smoothness of the guide walls and also on the material from which it is fabricated. Generally, the waveguide is carefully drawn from billets of copper, brass, or aluminum. The loss for copper is less than 2 dB per 100 ft at 4.2 GHz. Pieces of waveguide or components fashioned from waveguide are joined together by carefully machined flanges on mating parts.

B. Round Waveguide

Round waveguide is also widely used for connections from the electronic equipment to the antenna system because of its lower loss. Round waveguide can support several modes, depending on the frequency and how the waves are launched into the guide. Being a symmetrical structure, two polarizations can be carried, and by careful design, it can be used for several frequency bands simultaneously. For example, WC-281 (2.81 in. in diameter) is used for systems in the 3.7- to 4.2-GHz, 5.925- to 6.425-GHz, and 10.7- to 11.7-GHz bands. Waveguide components that are complicated in mechanical design or are difficult to manufacture by casting or machining techniques are often electroplated using precision mandrels to control the size and shape of the part.

C. Elliptical Waveguide

Waveguide with an elliptical cross section is often made from an intricately formed mental ribbon that is wound on a shaped mandrel in such a fashion that the finished product is semiflexible and can be bent on a large radius without deforming the cross section of the guide. This structure will support only a single polarization in the TE_{01} mode. Its use has been limited and is chosen primarily for its lower cost and the ability to manufacture it in long sections and to curl it up in large coils for shipping. Field flanging techniques have been developed and installation is relatively simple.

D. Square Waveguide

Square waveguide is rarely used but has some applications in waveguide components that require the propagation of two polarizations at the same time and for which circular guide is unsuited. An example is a case in which two polarizations of a signal from the round antenna waveguide must be converted into two rectangular waveguides at the equipment location.

VI. ANTENNAS

One of the most important components of a microwave system is the antenna. It is important because (1) system performance is highly dependent on its characteristics;

(2) the antenna and its supporting structure (tower) are physically the largest part of the station; and (3) the antenna tower structure and the ancillary components, such as the antenna feed line and supporting hardware, represent a sizable portion of the cost of the repeater installation. The microwave antenna is physically large because it must provide high gain to make up for the huge propagation loss that exists between the two repeater sites. The loss between two isotropic antennas operating on a frequency F (in gigahertz) with a spacing of S (miles) between the sites is given by the following:

$$\text{loss} = 20 \log 71{,}000 \times F \times S.$$

At 4 GHz with $S = 25$ miles, the loss is 137 dB.

Antenna design, predicated on the principles of reflection and refraction of electromagnetic waves, has led to three basic types of structure. The first and most widely used is the parabolic reflector or some portions of a parabola, followed by the lens and then by the passive reflector or periscope. Two types of lens antennas have been used, namely the accelerator lens and the delay lens, each acting on the emerging wavefront. These designs are inherently frequency sensitive and therefore cannot be used in multiband systems. The New York–Boston route used the accelerator-type lens, while the early transcontinental 4-GHz stations employed the delay lens structure. Neither of these antennas is used to any extent today because the parabolic reflector design operates effectively over wider bandwidths and with two degrees of polarizations.

A. Parabolic Reflector Antenna

A parabolic reflector, with an antenna feed at the focal point, will radiate energy from the reflector with parallel rays that can be focused. Similarly, parallel rays entering a receiving antenna reflector will concentrate the energy at the focal point. In the case of a transmitting antenna, a beam is formed which can be concentrated and directed toward a distant target, and theoretically, no energy is transmitted backward from the reflector. Similarly, in the case of a receiving antenna, little energy will be received from the backward direction. A microwave antenna is stated to have gain compared to an isotropic radiator; yet it is a passive device and no external power is added. This so-called gain is achieved by the high directivity obtained from the large diameter of the reflector compared to the wavelength of the signal and the use of a signal launcher that illuminates the reflector from the focal point. The larger the area of the parabolic surface, the greater the directivity and, hence, the gain. For example, an antenna 3 m in diameter has a gain at 4.2 GHz of approximately 40 dB. One watt of power from the antenna feed at the focal point will provide an effective radiated power of 10,000 W and will have a beam width of approximately 3° ($\pm 1.5^\circ$) at the 3-dB down point of the major lobe. This requires high accuracy in aiming the antenna toward a target 25 or so miles away. The gain of the antenna in decibels is equal to 20 log $7.8FD$, where F is the frequency in gigahertz and D is the diameter in meters referred to an isotropic radiator.

In terrestrial microwave systems, the antenna is generally pointed at or near the horizon, and for a parabolic antenna on the top of a tall building where the SHF equipment may be nearby, rectangular waveguides can be used to connect to the antenna feed elements. Two polarizations will require two waveguide runs. If the antenna is on a tall tower and the equipment is at ground level, the two rectangular guides needed may have excessive losses. It is difficult to connect a direct radiator parabola with circular guide because of the vertical-to-horizontal transition needed at the top of the tower. The horn reflector antenna, invented by H. T. Friis in the mid-1940s and in wide use today, uses only a section of a very large parabolic surface as the radiator and is illuminated from the focal point by a tapered horn of either round or square cross section in such a way that the vertical-to-horizontal transition takes place within the antenna itself. The horn has a carefully shaped transition taper from round to square to connect the WR-281 circular waveguide to the square throat of the antenna with minimum mode conversion effects. This use of circular guide allows the microwave equipment to be located at the base of the tower without incurring prohibitively high losses. This antenna has metal sides for structural reasons and, therefore, has excellent side-to-side shielding, permitting several antennas to be mounted on the same tower platform. It has excellent radiation patterns, high gain, and a large front-to-back loss ratio. This antenna can be identified in the countryside by its appearance, which resembles a cornucopia or a large ice-cream cone atop the tower, and frequently there will be a cluster of four or more horn reflector antennas at large repeater stations. In recent years, this antenna, made by a number of vendors, has been altered to include a conical cross section from the circular guide to the parabolic surface.

B. Receive Television Only

The parabolic antenna is gaining wide acceptance for receive-only earth stations to provide home reception from television satellites. Many households today display a dish in the yard, which is not easy to hide and still have it function efficiently. It may become a status symbol despite its unaesthetic appearance in some neighborhoods. Most home antennas are $\sim$3 m in diameter and, as previously stated, have about 40 dB of gain, which usually provides an acceptable picture for home viewing.

Satellites for commercial applications such as telephone, data, and network television transmission use much larger parabolic antennas; some are as large as 10 m in diameter and have gains approaching 60 dB at 6 GHz. This is necessary to provide the performance needed for these services. Harvard University has a radio astronomy parabolic antenna 25.6 m (84 ft) in diameter with a gain of 58 dB at the frequency of interest, which is 1.25 GHz. At Arecibo, Puerto Rico, the world's largest antenna is 1000 ft in diameter and has 60 dB of gain at 430 MHz. In some areas of the country, large dish antennas must be heated to prevent ice and snow from collecting on the parabolic surface and distorting the wave pattern, which would otherwise introduce excessive losses.

C. Passive Reflector

The last type of antenna that has been mentioned, and which was in widespread use in the 1960s for short haul, is the passive reflector. In this case, the usual configuration is a parabolic antenna mounted at ground level and radiating vertically. A nearby tower supports a large metallic plate, usually at about 45° with respect to the ground, which serves as the radiating element. This arrangement saves the expense of the waveguide feed line up the tower, but the reflector is difficult to adjust because two angles are involved (the angle of incidence and the angle of reflection), and it also has the tendency to allow energy to spill past the reflector and interfere with other services, including orbiting satellites. This antenna has been "outlawed" for new construction, although many still exist under a grandfather clause.

VII. COMPONENTS

Many components go into a transmitter–receiver unit and the ancillary subsystem parts needed to make up a total transmission facility. The evolution of microwave components from the days of New York to Boston to the present day is a saga of the work of thousands of individuals and groups that honed the technology from a crude beginning to a high degree of sophistication for a wide variety of communication uses and needs. As new technology, materials, processes, and inventions come into being, microwave system designers keep a careful eye on how these can be used to improve or expand the capability of the systems at hand or how they can be used for the next system to be developed. Such scrutiny in the past has paid huge dividends, and today's advanced designs are far, far ahead of those of a generation ago. The array of components and subsystems that make up the transmitting and receiving portions of a microwave system and the integration of this equipment into the repeater installation requires meticulous attention in the design and the manufacture of each of the individual parts.

A. Klystrons

The source of microwave energy for the New York–Boston system, and many other early systems, was the klystron electron tube, a device that, because of its internal construction, can be made to amplify or oscillate at SHF. These tubes were invented prior to World War II and became the first available for either laboratory experimentation or working systems. A reflex klystron has a resonant cavity that can be tuned over a small frequency range by exerting pressure on the cavity wall to change the internal dimensions of the tube and, therefore, the frequency of oscillation. Multicavity klystrons can be used as amplifiers by injecting an input signal into one cavity near the gun structure and extracting the amplified signal from another cavity further along the electron stream. A two-cavity klystron was used in the transmitter of the New York–Boston system.

B. Planar Triode

It was learned even before the New York–Boston experiment that the klystron tube would not be satisfactory for long systems, and in the mid-1940s work had started on a three-element electron tube that would amplify at 4.2 GHz. J. A. Morton and others conceived a high-performance vacuum tube that had internal electrode sizes and spacings only a fraction of what had been used before in other electron devices. The tube that evolved had a grid structure whose wire was only 0.0003 in. in diameter and which was tension wound 1000 turns to the inch. The spacing between the grid and the cathode is 0.0005 in., which is much less than the diameter of a human hair. The transconductance of this triode is 50,000 μmho. This tube is known as a planar triode tube because the grid is a flat, meshed disk sandwiched between a flat cathode and a flat plate. Operation of this early microwave triode in a grounded-grid, tuned cavity structure had a power output of only 0.5 W across the 3.7–4.2 band. The power limitation was one of poor heat dissipation of the glass used as insulation in the tube envelope. The input power to the tube had to be kept low enough to avoid damage and early failure of the device.

This tube has turned out to be an extremely important component for one very large system, and development efforts to improve its performance, life, and output power and reduce its cost continued for 25 years. Replacement of the glass insulation with beryllia (a ceramic with excellent insulating and thermal conductivity properties) in

the late 1960s, in part, contributed to an increase in the safe operating power level of the tube. Improved versions have provided 5 W of output power in some applications. Today, this tube has been replaced by solid-state gallium arsenide field-effect transistor (GaAs FET) amplifiers in virtually all applications.

C. Traveling-Wave Tube

About the time that Morton started work on the planar triode, Rudolph Kompfner of England conceived a device that became known as the traveling-wave tube (TWT), which consisted of a gun structure to focus an electron beam down the center of a small-bore helical winding and a collector to receive the high-speed electrons at the far end. The microwave signal to be amplified is coupled into the helix at the gun end, and the electron beam is focused along the tube length by a precision magnetic field around the outside of the tube envelope. The signal propagates down the helix, gaining energy from the electron beam. Output is coupled out of the tube near the collector. Traveling-wave tubes for transmitter applications have been made with power outputs ranging from a few watts to several hundred watts in the SHF band. One virtue of the TWT is its very large bandwidth.

The first application of a TWT amplifier in a long-haul system was in a 6-GHz application in 1960, and since that time a number of other applications have used this basic design. Unlike the planar triode, where the life is 15,000 hr or less, the TWT life is generally on the order of 100,000 hr or more, approaching what can be achieved with solid-state devices. The cost of a TWT is at least 10 times that of the triode, but the TWT can operate at much higher frequencies and at higher power levels. One disadvantage of the TWT is that for 10 W of output, the power supply voltages are on the order of 3000 V, necessitating elaborate safety precautions to protect maintenance personnel.

Other components for microwave systems include high-frequency diodes, rf converters (mixers), wave filters, equalizers, isolators, circulators, and generators of microwave power. Over the years, many other systems have evolved in other frequency bands, and in each case the components have been tailored for the specific applications. All, however, require components and subsystems to perform essentially the same functions.

D. Microwave Networks

Wave filters are needed to select specific frequencies and to reject others. In low-frequency communication applications, frequency selectivity is achieved by the use of discrete capacitors and inductors arranged in resonant circuits. At SHF, these discrete components cannot be used effectively; rather, distributed elements, usually in waveguide or strip-line structures, must be employed. In a waveguide filter, irises in the guide and posts across the guide perform the resonant functions that coils and capacitors achieve at lower frequencies. Over the years, an entire technology has been developed for ready computation of various characteristics of waveguide networks. Originally, these involved computations that were performed on slide rules and mechanical calculators requiring a substantial staff of people spending many hours in the design process, but today the electronic computer and the many programs available have reduced this design effort to routines rapidly carried out by relatively nontechnical personnel.

Recently, a new material known as yttrium–indium–garnet (YIG) has become available which has ferroresonant properties with high stability that are almost ideal for microwave filters. Its great virtue is that filters made from this material are very much smaller and lower in cost than conventional filter structures. Both of these factors are highly significant.

E. Ferrites

About 1960, a new material called *ferrite* became available, which has proven to be highly useful in a number of microwave components. Ferrites can vary in composition, depending on the specific results desired, but most contain a mixture of the oxides of nickel, iron, copper, and aluminum. Physically, ferrites are a form of ceramic for they are hard and brittle and must be either molded in the green state and kiln fired or ground to size after firing. These materials, when used in a waveguide structure with associated magnets, can form a nonreciprocal (one-way) transmission path; that is, microwave energy can propagate in one direction through the device, but energy flowing in the reverse direction is absorbed in the material.

F. Isolators

Depending on the composition of the ferrite material and the shape of the piece, it can exhibit either field displacement or Faraday rotation characteristics. The field displacement material is often used for isolators, which are made by mounting the ferrite slab lengthwise in the guide and perpendicular to the wide walls. An external magnet coupled into the edge of the slab provides the polarization field necessary to achieve directivity.

G. Circulators

Circulators use the Faraday rotation type of material with the ferrite mounted in the center of three converging

microwave paths, each 120° to the adjacent path. The ferrite in a circulator is usually in the form of a small disk with the external magnet directly over the area of the ferrite. Depending on the polarity of the magnet, the energy can be circulated from one of the three ports to an adjacent port either to the right or left. If circulated to the right, energy cannot go to the left port. If the polarity of the magnet were reversed, the circulation would be in the opposite direction. The circulator provides great flexibility in the design of microwave integrated circuits.

H. Downconverters

In the 1950s, the conversion from SHF to the 70-MHz IF was accomplished with a point-contact silicon diode and a microwave source as a local oscillator (at that time the oscillator was a reflex klystron). This mixer or converter, as it is sometimes called, has loss, and the point-contact diode introduces noise which can result in a receiver noise figure over 12 dB. In the late 1960s, the Schottky or hot-carrier diode became available, and because of the way the junction is formed, it is an ideal mixer at SHF. A 4-GHz Schottky diode and associated structure can have a noise figure of less than 8 dB. This was a major step forward, and in one case, it resulted in a very sizable increase in the number of voice and data channels that could be carried over a single radio channel. The Schottky diode is rugged, it can sustain substantial overload from the local oscillator and the signal, and it can be manufactured with relative ease.

I. Tunnel Diode Amplifier

Solid-state diodes have been used to provide amplification by using the negative resistance property of the device. The details of diode characteristics, how doping in the course of manufacture can alter their properties, and how negative resistance can occur can be found elsewhere. By embedding a specially doped diode in a strip-line circuit and applying an appropriate bias to the diode, an amplifier can be designed. One type is the tunnel diode amplifier (TDA), so named because of the special fabrication of the semiconductor junction, which can provide low-level gains up to 20 dB with noise figures as low as 4 dB. The bias voltage on the diode to optimize the negative-resistance characteristic is very low, on the order of 0.4 V, and to avoid swinging outside this region, the input signal must also be kept low. Due to this tendency to readily overload, the TDA has found little use in terrestrial systems even though its gain and noise figure are quite good and the design is not very complicated. Tunnel diodes have been used effectively in some phase-array radar systems primarily because the radar return signal is usually not high enough to cause overload difficulties. Germanium diodes are frequently used in TDAs at frequencies up to almost 30 GHz.

J. Parametric Amplifiers

Another type of negative-resistance amplifier for receiver applications which is more complicated than the TDA but which does not suffer from the low-level limitation of the TDA is the parametric amplifier. This type of circuit requires that the diode be "pumped" or driven by a local oscillator source at several times the operating frequency. For example, a 4-GHz amplifier is pumped by a 12-GHz source that must be very carefully stabilized in amplitude.

Parametric amplifiers were thoroughly evaluated for widespread use in the long-haul environment in 1968, but it was found that the circuit complexity needed to keep the pump signal at the optimum value was impractical and unduly expensive. This amplifier is no longer in use in long-haul terrestrial common-carrier systems, although they have been used in some satellite applications.

K. IMPATT Amplifiers

Impact avalanche transit time (IMPATT) solid-state power amplifiers also use the negative resistance of the diode in specially designed circuits that include circulators and strip line to connect the devices together. Both silicon and gallium arsenide materials have been used, depending upon the power level at which the amplifiers are to operate. In relatively low-power amplifiers (1–2 W), a single diode can provide the required output, but for higher powered applications, several devices must be cascaded.

IMPATT amplifiers are basically reflection amplifiers. The input signal from one port of a circulator is terminated with an IMPATT diode that is suitably biased into the negative-resistance region. Amplification takes place because negative resistance has been introduced into the circuit which cancels out the losses so that the output signal is higher than the input. The output is returned to the same port, circulated to an adjacent port, and then delivered to the system output in the case of a single-diode unit or to the next stage of a multistage amplifier. The output cannot be conducted toward the input because of the high reverse loss of the circulator.

In the case of single-diode amplifiers, approximately 20 dB of gain can be realized at frequencies in the 6- to 11-GHz range. Tuning procedures, while not complicated, frequently require caution by the operator to avoid improper matching of the diode into the circuit. If this should happen, there is a possibility that the diode will be permanently damaged. Electron tubes are much

more tolerant of misadjustment than IMPATT diodes. Noise figures of IMPATT amplifiers are generally high; some typical power amplifier noise figures range from 30 to 50 dB, which is higher than those for electron devices.

In multistage amplifiers, each diode must be individually biased and it is necessary to carefully regulate the voltage to each semiconductor. The IMPATT amplifier can be highly nonlinear, and therefore it is not well suited for AM systems. These nonlinearities do not degrade the performance of angle-modulated systems using either FM or PM. In some applications requiring high gain, the IMPATT diode may function as a locked oscillator, where the input signal causes the oscillator to instantaneously follow the input frequency exactly. This is the equivalent of an amplifier, for the output signal is a replica of the input but the amplitude is the equivalent of that of the oscillator. Precautions must be taken to assure that in case of failure of the input source the amplifier is adequately squelched to prevent interference.

The IMPATT amplifier has been successfully used in 6- and 11-GHz systems operating in the 1- to 3.5-W power range with quite satisfactory results. These have been single-diode units of both silicon and gallium arsenide. As the output power level requirements rise, the maintenance difficulties, circuit complications, and costs accelerate rapidly. A five-stage amplifier requires five voltage regulators, output monitors on each of the five stages, an intricate strip-line circulator structure, and meticulous circuit adjustment. The power efficiency, that is, the rf output power versus the overall dc input power, is considerably lower for the IMPATT amplifier compared to the TWT. In large installations where many amplifiers are used, these power differences may be significant.

The component costs of the 10-W IMPATT amplifier, which can provide the same gain as a TWT (about 32 dB), are not substantially lower than the TWT and the high-voltage power supply that it was to replace, and while it has been demonstrated that solid-state 10-W amplifiers are feasible, they are not used in long-haul terrestrial common-carrier systems today.

L. Gunn Diode Amplifiers

The Gunn diode has been used in low-power microwave systems, often with the diode as a locked oscillator, as described above for the low-power IMPATT. The Gunn diode has lower noise than the IMPATT and is ideally suited to low-power applications (up to about 2 W). The Gunn diode can also serve as the receiver local oscillator when suitably stabilized in frequency from an external source.

M. Transistor Amplifiers

In the early 1970s, sufficient exploratory work had been completed on microwave transistors to indicate clearly that the GaAs FET offered intriguing possibilities in power amplifier, receiving amplifier, and frequency converter applications. The performance of this class of transistor was a result of the high electron mobility of the class III–V elements in the periodic table and the ability to design more nearly planar transistor geometries. These transistors had less stringent spacing requirements than the equivalent bipolar units and, as a result, could offer better high-frequency performance for a given degree of manufacturing complexity.

Circuit development ensued, and for the first time, active microwave circuits were being developed using an equivalent-circuit design approach that was sufficiently detailed to accurately predict performance and, at the same time, simple enough to lend itself to direct calculations of the device geometry. Amplifier circuits can be designed that cover 10% bandwidths, and therefore frequency adjustment is not required in the field. This is a distinct advantage over the application of the planar triode tube and the IMPATT diode amplifiers. A 2-W, 4-GHz amplifier was designed to replace the planar triode amplifier in many applications, and this was followed by a 5-W version that offered the possibility of increasing the capacity of older 2-W relay systems. Field experience with GaAs FET power amplifiers has indicated that after a short burn-in period failures are rare and performance is sufficiently stable that routine maintenance is not required. This results in great economic savings to those who own and operate these systems.

The low inherent noise of the GaAs FET transistor is used to increase the sensitivity of the receiver section of the radio repeater. Amplifiers with a noise figure of 2 dB have been built and provide an improvement by a factor of at least 2 in receiver sensitivity. To obtain the best noise figure, the gate lengths and widths in the transistor have to be kept to a minimum. The use of a common transistor low-noise amplifier in the rectangular waveguide run to a number of repeaters in tandem provides a low noise figure to all units at a considerable cost savings. A common amplifier in the receiver input, if not designed to have low intermodulation characteristics, can result in undesirable interchannel cross talk during heavy selective fading conditions. Therefore, linear amplifier designs are sometimes necessary in low-level applications in the receiver input.

In the early 1980s, further semiconductor developments had a continuing impact on the capability of components for microwave radio relay systems. The development of the field-effect transistor with two gate electrodes made possible the design of a new group of frequency

converters and gain control circuits with sufficient gain to mask the noise from subsequent circuit elements. Special doping formulations, the introduction of indium, and improved semiconductor geometry gave rise to higher electron mobility, closer spacings, thinner conductors, and better grounds. The ability to add circuit-matching elements on the semiconductor chip has increased microwave performance. Monolithic circuits with one or more stages on a single chip and high-electron-mobility devices are examples of these improvements. Circuit developers have been able to introduce meaningful computeraided design programs that enable more complete evaluation of circuit and environmental options in the design stage.

Typical performance capabilities now include power levels as high as 100 W at low microwave frequencies (1 GHz) and as has high as 8 W at 18 GHz. Bandwidths with relatively constant gain have been extended to an octave or more. Amplifiers for quadrature amplitude modulation digital modulation have power outputs of 4 W or more at saturation and gains and linearities that equal or exceed those of TWTs. Low-noise amplifiers have noise figures of 1 dB at 1 GHz, 1.3 dB at 6 GHz, and 3 dB at 20 GHz. By refrigerating to 20 K, noise figures as low as 0.1 dB at 1 GHz, 0.25 dB at 6 GHz, and 0.9 dB at 20 GHz are possible. Using high-electron-mobility transistors, a room-temperature noise figure of 1.4 dB at 11 GHz has been obtained. If refrigeration is used, the noise figure drops to 0.35 dB.

Circuits that combine a dielectric resonator and a GaAs FET transistor in an oscillator configuration can provide a stable frequency source with a stability of a fraction of a part per million per degree Celcius, which would satisfy short-haul requirements. For more critical operations, a transistor operating in a phase-locked loop with a low-frequency reference offers even better frequency stability.

N. Strip Line

The invention of the transistor, which is considerably smaller than the electron tube, required new methods of mounting it and the other electronic components needed to form a complete circuit. For many years, printed circuits have been routinely employed where the electrical and thermal demands are not great and where the use of phenolic or epoxy glass board does not affect the circuit performance. At SHF, these materials are not thermally stable and dielectric losses can be excessive. An improved technology is available for these high frequencies called *strip line*, which is a process that deposits a conductive path on a high-quality ceramic substrate. The pattern is usually of copper or gold, which provide high conductivity to keep the circuit losses low. The pattern is actually the electrical circuit with inductors, capacitors, resistors, and the interconnection paths formed by the deposited film on the substrate. At these frequencies, inductors and capacitors are physically small, and in a strip-line design an inductor may be only a narrow strip of metal on an otherwise plain section of the ceramic sheet. Intricate circuits are possible; the semiconductor active components can be appliquéd onto the circuit. By this means, the leads can be kept very short, which is a necessity at these frequencies. Manufacturing reproducibility is very good with strip line because the metal deposition onto the ceramic is extremely accurate with the photographic techniques used.

Strip line is used for many SHF circuits in transmitters and receivers. By integrating stripline circuits with ferrite circuits, very elaborate microwave subsystems can be devised. For example, the entire SHF portion of a transmitter or receiver can be designed as a single unit.

VIII. RELIABILITY

Generally four factors can impair the reliability of a microwave system, and each requires specific action to minimize its impact on service quality. These four causes of degradation are (1) propagation effects, (2) equipment failure (3) power failure, and (4) maintenance errors and/or personnel intervention. All of these factors have been minimized by additions and improvements to the systems and by adequate personnel training in the operation and maintenance of the facilities.

In the case of multipath fading or equipment failures described in the next two sections, circuit outage can be greatly overcome by the rapid substitution of a good channel for the one that has failed. Protection switching is described in Section VIII.E.

A. Propagation

Long ago it was observed that the input signal from the antenna into the receiver was not always constant. When more information was obtained from various parts of the country, it showed that some sections were more prone to fading than others; the months of July and August were worse than the other months from the standpoint of the severity and frequency of occurrence, and the time of day had an effect on the constancy of propagation. The flat plains areas, particularly in the time period from midnight until after sunrise, had more fading than the irregular topography of the Rocky Mountain region. Reflections from flat ground with peculiar temperature conditions and still air enhance the fading. The locale around the Gulf Coast area, Florida, Georgia, and Alabama is also a high-fade region that is caused by atmospheric conditions.

Propagation phenomena have had, and continue to receive, a large amount of theoretical study and field verification. Crawford and Jakes in 1952 reported on selective fading at SHFs and DeLange made tests at these frequencies using very short pulses. Kaylor in 1950 studied fading at 4 GHz by very elaborately instrumenting (for that period) one 30.8-mi hop between Lowden and Princeton, Iowa. These tests showed some remarkable effects that led to corrective steps to minimize the outages due to fading. This is covered is Section VIII.E on protection switching.

Fading can be of two types: a partial or total obstruction of the path that might exist for some period of time (from minutes to hours). This has the same effect as if a "bulge" in the earth had risen up to produce a path blockage. Earth bulge fading affects all channels at the same time for it has no frequency dependence (flat fade). Fortunately, this kind of fading is relatively rare in both location and frequency of occurrence.

Multipath or selective fading occurs when the atmospheric conditions cause the various rays of the radio wave to be reflected and/or refracted over the path until some arrive at the receiving antenna out of phase with the main or direct wave. The cancellations at the receiver, either in total or in part, affect the received signal level. This partial cancellation in a single 20-MHz channel can be highly selective; it has frequently been observed that a fade can be much more severe at one part of a 20-MHz channel than at another. This leads to serious distortions of the signal and often makes the channel unusable. At other times, the fade will affect the entire channel about equally and the channel will become noisy. This is frequently observed on a television broadcast that is being carried by terrestrial radio facilities. Selective fading is very rapid, often measured in terms of a few seconds per occurrence. A great wealth of published material on atmospheric fading at SHF exists for the interested reader.

It has already been mentioned that at frequencies above about 10 GHz rain affects the transmission as though the transmitter power had been lowered or an attenuator had been inserted into the input of the receiver. This rain attenuation is caused by the dispersion of the radio wave by the individual rain drops as they fall to earth. In normal or dry circumstances, the wave is directed to the target (receiver) by the transmitting antenna and the path, so established, is then considered the norm. When falling rain is interspersed over the path, the rain drops act as small refractors and divert the beam off the norm. More rain, from either higher precipitation or longer paths (more rain drops), will produce higher attenuation. Rain effects are frequency sensitive only to the extent that all frequencies above about 10 GHz will be affected, and the higher the frequency the less rainfall is needed to produce a given amount of attenuation. Similarly, the greater the rainfall, the higher the attenuation, and the longer the path over which the rainfall exists, the greater the attenuation. For this reason, the 11-GHz band is usually used in short-haul systems only, and then only when rain effects will not be serious. However, the 11-GHz band is not a total loss even in high-rain areas because acceptable performance can always be achieved by reducing the distance between the repeaters. By doing this, the cost of the system will increase due to the additional sites and equipment that are needed. Instead of the 26-mile average spacing that prevails at 4 and 6 GHz, the 11-GHz systems average about 15 miles between repeaters. At frequencies in the region of 18 to 20 GHz, the repeater spacings may average only 2–4 miles because rain attenuation will exist at more locations and at lower levels of precipitation.

B. Equipment Failure

Microwave equipment, like all electrical and electronic equipment, will ultimately fail, and when this occurs, the remainder of the system will not be usable until repairs are made or the system is restored with an automatic protection system. Over the years, every subsystem, equipment unit, and component has been studied and the industry has spent great effort and expense in redesign and in introducing new technology to make the individual parts of a system more reliable. The results have been very effective, and the life of the components, including the electron tube devices, has been greatly extended over the years. Manufacturers of components have also made improvements by studying their processes and making changes whenever needed, and tests have been added at the time of manufacture to assure that early failure in the field will not occur. Only two communication facilities are given more attention and scrutiny in design and manufacture than microwave relay: they are (1) undersea cable systems and (2) the orbiting equipment units of satellite systems.

Since solid-state components have become available, many circuits of existing systems have been redesigned to take advantage of both improved performance and higher reliability. Of the latter, it can be said that very often a redesigned circuit with solid-state devices can pay for itself in the reduction of maintenance costs in a short while. For example, in one large system that had used the planar triode tube, redesign to introduce the GaAs FET power amplifier has replaced virtually all of the 135,000 tubes (system wide), which had required maintenance at least once a year. Considering the locations of all the repeaters throughout the United States, a great amount of time was required to get to the site, make the repairs, and return to the home location. Studies show that the average maintenance call from a center to a remote repeater location is 1.5 hr. New designs, therefore, are not always used for

the manufacture of equipment of new systems; frequently, they are retrofitted into existing systems to save money.

C. Power

Home electronics equipment usually operates from commercial alternating current, and when a power failure occurs, the other uses of electricity are also affected. Generally, the user associates the loss of television or radio reception with the local power failure, and when the lights come back on, usually the television is again operating normally. The home consumer accepts this situation as an "unavoidable" event, even though an automobile crashing into a utility pole has severed the power line. When the same conditions occur a hundred miles away and telephone, data, and television circuits are interrupted, the facility users are less tolerant of a power failure and they demand that the carrier do better. To overcome system failure of this type, the carriers generally have established emergency power facilities at each repeater site. Battery standby is by far the simplest and cheapest, and on light route systems, this may be the only backup. Capacity is frequently set for 8 to 24 hr with the expectation that the commercial power will be restored by that time. Oftentimes means are provided for bringing in a portable engine generator set, towed by a truck, to connect into the station to recharge the batteries if it appears that the loss of alternating current will be prolonged.

In the larger stations, the battery backup and the engine alternator are a part of the permanent station installation and automatic startup of the engine is provided. With such an arrangement, there is no interruption at all in the continuity of service. With added control and command facilities at each location, maintenance people can periodically "exercise" the reserve power system remotely to be assured that the equipment is in proper working order.

D. Maintenance Errors

In large systems, such as a transcontinental route, many service control points are involved and a large number of people are charged with the maintenance of the system. It has been learned from experience that even a single operating error can effect a serious outage; a single erroneous push of a switch button can cross CBS with NBC. Operating errors can happen, but the carriers have laid great emphasis on personnel training and the importance of avoiding operating errors. In general, these people are well informed and perform at a high level of competence. By making the equipment more reliable and by making many of the functions automatic, people are less involved and the number of manmade interruptions are reduced accordingly.

E. Protection Switching

Fading of the microwave signal can produce interruptions in service from momentary outages to prolonged intervals when the radio channels are unusable. Long outages are usually the result of earth bulge atmospheric effects and almost always degrade all channels simultaneously. The other type of fading is caused by multipath transmission, and on any specific channel the disturbance is only momentary. All channels in a system might experience serious fading in the same time period but not at the same instant. Equipment failures can also cause circuit outage, but with modern-day equipment, this is no longer the problem that it once was with short-lived components. Unfortunately, specific equipment failures cannot be anticipated and such events always cause an interruption in service.

For both multipath fading and equipment failures, microwave systems have rather sophisticated schemes for using one of several authorized channels to provide protection for the remaining ones in the group (as many as 11 channels). This channel substitution (channel switching) is done very quickly, and even with unanticipated equipment failures, service is restored in less than 50 msec. Protection arrangements of this type are ineffective for earth bulge fades because one protection channel cannot protect a multiplicity of channels that have failed simultaneously and in which there is a high probability that the protection channel in the same band would also have failed under these circumstances.

Protection switching in long-haul systems is done by sectioning the route into what is known as a switching section, and each segment is operated independently of the other sections for protection purposes. Telemetry functions are provided between the ends of the section to control the switches. In practice, a switching section may be a single hop or it may be as long as about 12 repeaters (longer sections become impractical for a number of reasons.) The average section length in the United States is three hops, which is predicated on the need to provide protection switching wherever the baseband signal is to be delivered to the user. Without this requirement, many longer switching sections would exist.

Channel switching is done at the IF with solid-state switches. A switching matrix is arranged to substitute a channel known to be good (the protection channel) for the failed (working) channel, and in a multichannel system this requires a switch at both the head end and the tail end of the section. The quality of each channel, including the protection channel, is monitored for noise and the presence of a pilot tone. When the detector at the tail end of the section senses an abnormality (high noise or lack of pilot or both); a switch is initiated by sending a control

signal to the head end (transmitting end) to make a bridge between the working channel and the protection channel. The monitor on the protection channel recognizes that the head-end switch has been made and proceeds to connect the output of the protection channel to the working channel. These precautions, namely verifying that the service is on the protection channel before the final switch is made, give assurance that a switch to a dead channel will not be completed.

The failed working channel is monitored continuously for noise and presence of pilot even though the channel is disconnected from the system. This is done to determine when the noise performance is such that the working channel can be returned to the system so that the protection channel can be released and made ready to serve some other channel that requires protection.

Current systems use microprocessors for logic control of the switches that can be used for more than the protection of failed channels. Switches can be made manually or automatically for maintenance checking and for circuit rearrangements either on demand or on a predetermined schedule, such as television feeds of a network program to a local outlet or a television pickup of some event of interest that is to be connected to a television network.

In short-haul systems where baseband signals are available at each repeater site, switching is usually done at the baseband frequencies. Depending on the number of working channels in the system, head–end switching at baseband may or may not be used. In light systems, two repeaters operating on the same frequency may be used, one being a hot standby for the other, with a simple transfer switch at the tail end to make the transfer when ordered by the switch-initiating circuits.

F. Space Diversity

The large amount of attention given to microwave propagation many years ago established the fact that if two receiving antennas are located on the same support but separated vertically by 35 ft or more, the output of the two antennas, under fading conditions, is almost totally uncorrelated. This says that at any given instant, if one antenna output is faded and unusable, the likelihood that the output of the other antenna is normal or quite acceptable is very high. Successful reception virtually 100% of the time then depends only on providing the means to select the right antenna at the right time. Over the years, many clever schemes have been devised for detecting and selecting the good signal. An arrangement for combining the outputs of both antennas, with appropriate phase shifters to account for the differences in the two signal path lengths, has been devised so that a good signal is available essentially all of the time. In many systems, such as single-sideband and digital systems, space-diversity arrangements are necessary.

IX. RADIO TERMINALS

The information-bearing signal that is to be transmitted via microwave facilities is always generated at a lower frequency and is then translated (modulated) up in one or more steps to a discrete frequency in the SHF band. This lower frequency is called the *baseband* and may extend from virtually direct current to ~8–10 MHz. Television baseband signals generally will be originated by the broadcaster and will contain the sound as well as the picture information. It is delivered to the carrier in a very standard format; usually the sync pulse information will be more negative than the black or blanking level, and the picture information will be positive relative to the blanking level. Voice-frequency telephony, on the other hand, will be delivered to the carrier as a broadband signal resembling random noise after the individual voice-frequency circuits have gone through several steps of translation to place each 4-kHz circuit in a discrete 4-kHz slot in the baseband spectrum. Digital signals may either be handled as baseband signals or by special signal processing to get a suitable IF format to be applied to the radio. More information on these terminals can be found in Section XIII on digital radio.

A. FM Radio Terminals

For FM microwave systems, the FM terminal equipment consists of an FM transmitter and an FM receiver. The FM transmitter contains an oscillator circuit that is capable of having the rest frequency deviated above and below this value by the baseband signal. The greater the amplitude of the baseband signal the greater the FM deviation from the center frequency. In recent years, voltage-controlled diode capacitors have been used in the frequency-determining circuits of the oscillator to produce the instantaneous changes in frequency, thereby generating frequency modulation. Over the years, considerable attention has been given to the means for controlling and stabilizing the average or rest frequency at the desired value. This is important because the IF signal from the FM transmitter is added directly to the channel local oscillator source to produce the SHF signal that is radiated from the antenna. If these signals are not accurately controlled, interferences in multichannel systems can be severe.

At the receiving end, the SHF signal of each radio channel is demodulated with the use of a discrete local oscillator frequency to produce the IF signal. This FM IF signal is then detected to baseband in an FM discriminator quite

similar to that used in commercial FM radio receivers except with much wider bandwidth. The baseband signal is then delivered to the end user, the television broadcaster, or the telephone company.

FM transmission consists of carrier and FM sidebands above and below the carrier and, as such, is a reasonably rugged signal format. Frequency shifts through repeaters and through the many stages of frequency translations that occur in long systems do not alter the frequency of the recovered baseband. This is because the discriminator in the FM receiver uses the carrier frequency to demodulate the FM sidebands. All frequencies are shifted the same amount in going through the system and any frequency shift is canceled out. This is not true in radio terminals used in SSB systems.

Baseband circuits in the terminals are used to adjust the levels of the signals and can provide means for equalizing the gain–frequency response characteristics of the various parts of the system. Because of the wide variety of signals that must be accommodated, the baseband circuits must be broadband.

B. AM Radio Terminals

Single-sideband AM radio systems are used to transmit information by using the spectrum in a more efficient manner than that of FM systems. Where FM equipment transmits information using two sidebands around the carrier that on the average require appreciable power for the carrier itself, SSB not only uses just one sideband but also the carrier is suppressed, thereby saving the power of the carrier. This higher efficiency in the usage of the spectrum is obtained at the expense of a very elaborate and precise control of all the frequencies in the system, including those used in the terminal equipment.

In telephone service applications, it is absolutely mandatory that virtually no frequency shift take place from end to end; otherwise, the voice quality is impaired and the customer will complain. Data systems produce errors and telephoto circuits produce distorted pictures if the frequency shift is more than 20 Hz. Frequency translation throughout the multiplexing hierarchy must be strictly controlled by locking to a high-precision standard. More information on frequency control can be found in the section on SSB AM radio.

X. SYSTEM TESTING

When microwave systems were first used for communication purposes, it was believed that a considerable amount of testing would be required to assure that reliable service could be provided at all times. It was felt that in a long chain of repeaters performance would not be stable, that frequent failures, especially of the electron tubes, would occur, and that the equipment would degrade with time. It was assumed that by frequently visiting the repeater sites and testing the electronic circuits, failures could be anticipated to some degree. Routine measurements for a number of key characteristics were made on many parts of the repeater with the expectation that they would essentially eliminate the equipment as a cause of trouble. These routine tests were made for many years and at considerable cost, and in retrospect, one might question whether this practice was cost effective. As equipment performance and reliability improved with a maturing technology with more reliable components, it became apparent that less and less maintenance was really necessary. Eventually, maintenance was done only on a demand basis (i.e., when something either failed completely and station alarms pinpointed the location or end-to-end performance had degraded beyond acceptable limits and terminal alarms called attention to the problem).

It should not concluded that the carriers no longer test systems, repeaters, and parts. Very elaborate test facilities are available to diagnose trouble and locate the defective elements, but tests are no longer performed on subunits of a repeater in a system that is in no apparent trouble.

In television systems, the carrier's "customer" is technically competent in that the people who take the television signal from the carrier are critical observers who have the means for determining the quality of the circuit. Therefore, imperfections in the television signal can result in a complaint from the broadcast people to the common-carrier operators. Television system testing is exacting and detailed. The broadcasters and the carriers use essentially the same kind of equipment and one can verify the observations of the other. The broadcaster uses a multiburst signal wherein a number of discrete sine-wave signals, all at equal amplitude, are transmitted sequentially and at the receiving end are displayed on an oscilloscope. By observing the height of the display of each of these tones, which appear as a series of vertical bars, the response characteristic can be evaluated. If the high-frequency bars are lower than the low-frequency bars, the high-frequency response rolls off and the picture will lose fine-grain detail. As an example, the picture of the shirt of a football referee standing directly in front of the camera will show large and distinct black and white stripes; as he moves out onto the field and away from the camera, the stripes lose their distinct pattern and smear into a gray smudge. Differential gain and phase, a measure of the amplitude and phase of a reference test signal as the input amplitude is varied, gives an indication of the stability of the color of the picture. Poor differential gain will cause the intensity of the color to change when the scene changes from light

to dark. Similarly, poor differential phase characteristics will actually cause the hue of the color to change with the intensity of the general picture content. Noise measurements are simple to perform but are important to assure the broadcaster that the picture being sent out over the air is not snowy or flecked with noise bursts.

A number of tests can be performed on a microwave system that carries telephony signals. When many voice circuits are multiplexed together, the total baseband spectrum has almost the same characteristics as an equal band of random or "white" noise. Tests made by using a random noise source can determine the amount of cross talk between voice circuits that is being experienced. The noise source, connected in place of the telephone circuits, is equipped with narrow band-rejection filters tuned to selected portions of the spectrum. At the receiving end, a sensitive detector is connected to the output of the corresponding bank of bandpass filters. Since no noise transmission exists at the selected frequencies at the sending end because of the reject filters, any detected output at these frequencies at the receiving end is caused by intermodulation in the various parts of the microwave system. This is a very powerful tool in determining the amount of intercircuit cross talk present.

It is also important that the level of all of the voice circuits of the radio channel be essentially the same in a microwave system used for telephony. To check this, a variable-frequency oscillator at the head end and an oscilloscope at the receiving end will show if the frequency characteristic is "flat" (i.e., signals at all frequencies are of the same amplitude). This measurement is made by automatically "sweeping" the frequency from the low end of the band to the high end. The operator can readily see the actual performance on the oscilloscope.

In data systems, the all-important characteristics are the error rate and the degree of intersymbol interference that exists. To check the error rate, a signal of known characteristics is transmitted over the facility and an error detector is connected to the receiver output. By observing the count on the detector, the quality of the system can be easily determined. A typical number frequently quoted is a bit error rate of one in a thousand bits (error rate of 1 in 10^3). The intersymbol interference can be evaluated by sending a random data pattern and allowing the detector (oscilloscope) to examine all values of the wave. The pattern so displayed will appear as a series of traces that ideally will all cross the zero axis at the same time and all rise to maximum at the same time. With distortion in the system, these conditions will not be met and what should be a wide-open "eye" will begin to show more and more closure as the distortion increases. Since the detector in the system must accurately determine whether a pulse is or is not present at the time of decision, an open "eye" is highly desirable if low error rates are to be achieved.

The specialized test equipment used in current test programs, both at terminal locations and at repeater sites, is not inconsequential in quantity, degree of sophistication, or cost. Furthermore, it is mandatory that the personnel maintaining large microwave systems be adequately trained to use this array of precision test equipment, to interpret the results of the measurements made, and to effect repairs when tests indicate that the system is no longer meeting specifications. In order to maintain long-haul systems to the degree needed to meet present day service requirements, a uniform set of procedures and practices must be provided to the maintenance forces assigned to these tasks. Such practices must be easily performed and understandable and provide unambiguous results.

XI. SINGLE-SIDEBAND AM RADIO

It was noted in a previous section that a decision was made in the late 1940s to use FM rather than AM for the first application of microwave radio. It was the consensus of the decision makers then that the distortions arising from the anomalies of the phase characteristics of the components and subsystems could be more readily handled than those that would be associated with the nonlinear amplitude characteristics of these same elements. It turned out to be a wise decision since SHF FM repeater designs are much less complex and less costly than AM repeater designs. In the search for ways of adding voice circuit capacity to the larger systems, attention was given in the late 1960s to reexamining the SSB question in light of the advancements in the field of component technology and in the technical understanding of propagation characteristics, system design, and circuit capabilities. An intensive study indicated that an AM design was feasible and also economical when fully loaded.

The major deterrent that kept the industry from going the AM route in the 1940s and the one characteristic that had to be overcome in the 1970s was that of amplitude linearity. Amplitude linearity in an electronic circuit is the degree to which the amplitude of the output signal is a replica of the amplitude of the input signal. If the output is not identical (except for scale) to the input, this is equivalent to an output that is identical plus a component that is equal to the difference or the distortion. The latter is not a usable component of the signal and is actually an undesired perturbation of the signal. The various circuits of a SSB AM repeater were studied to see where nonlinearities existed and what might be done to reduce them. These undesirable characteristics are most predominant in those circuits that must deliver substantial amounts

of power, and in the case of a microwave repeater, this is the transmitter output device. In 6-GHz systems where the output power must be on the order of 10 W, the TWT is the logical choice for the output stage. In studying the TWT from a linearity standpoint, some internal design changes in the tube were made to improve the linearity, but this was still not sufficient to meet long-haul system requirements. A number of circuit innovations were studied and tested before a decision was made to employ a technique of predistortion to solve the nonlinearity problem.

Predistortion, as the word implies, is a means of altering the signal into the TWT in a predictable fashion so that this input, when passed through the TWT amplifier, will emerge at a higher level but with the undesired distortion canceled out. Such a scheme is not new per se, but it has not been used previously at SHF. It is necessary that components, including the TWT, be very uniform in their characteristics so that replacement tubes and other parts can fit the predistortion circuitry with only slight adjustment and so that a single predistorter will fit all equipment units.

All SSB circuits must be designed with this one characteristic at the top of the requirements list. The IF amplifiers (preamplifiers and main IF amplifiers) have to provide sufficient gain to boost the low received level to that required for the upconverter and to accommodate an atmospheric fade of approximately 40 dB. Automatic gain control (AGC) to compensate for this depth of fade must be provided and at the same time must not introduce amplitude distortion. The high degree of linearity required in SSB AM is one of the costs of getting substantially greater system capacity.

Most of the circuits of the SSB repeater are quite similar to their counterparts in FM systems, but because the SSB signal does not contain a rugged carrier within its format, the SSB signal can suffer distortion from selective fading more readily than FM. Equalization of the amplitude–frequency response characteristic is achieved by pilot-controlled regulators that increase or decrease the amplification at certain frequencies relative to the other frequencies in the channel spectrum. More precise equalization of the signal is another cost in getting the benefits of SSB's higher capacity.

The SHF signal that is radiated from the antenna at each repeater is derived by modulating the IF with a precise carrier frequency. In FM systems, this carrier is obtained from a stable crystal oscillator and harmonic generating equipment that is independent of equipment outside the repeater site. Frequency instabilities or offsets from the norm, while small, do not affect the demodulated baseband signal because of the nature of the FM receiver as mentioned in the section on radio terminals. These same frequency instabilities and departures from the exact design values would be intolerable in SSB AM systems. To get these oscillators at precisely the right frequencies, synchronization to a nationwide frequency standard is necessary and this standard must be available at every repeater location. This may mean that the frequency standard must be provided separately by wire or cable facilities or be locally derived from the difference between two precision pilots transmitted over the radio. To get an appreciation for the frequency accuracy required, a 20-Hz frequency error in a voice channel would be noticeable by almost every listener. Other services may be even more demanding because some customers may depend on the telephone system to transmit accurate signals for medical or scientific applications. Therefore, it is probably necessary to have an accuracy of $\sim$1–2 Hz from end to end. The SHF signal at 6 GHz is 6 billion Hz, and 1 Hz is an accuracy of 6×10^{10}. This frequency control requirement is another cost attributable to the use of SSB AM.

The disbursive medium, caused by selective fading, requires that in the areas where this phenomenon is severe, space-diversity antennas must be used to ease the amplitude equalization problem. This is still another cost that must be borne for the use of SSB AM. Despite the costs and added complexity of (1) overcoming component nonlinearities, (2) providing more complex amplitude equalization, (3) obtaining increased precision of beating oscillator signals at each repeater point, and (4) in some locations providing space-diversity antennas, the overall economics of SSB systems are still favorable and allow the maximum usage of the assigned frequency spectra.

XII. SATELLITE COMMUNICATION

Satellite radio transmission plays a very important role in providing international communication facilities (overseas circuits) that, at the present time, can otherwise be provided only by ocean cable systems. The latter have limited capacity, are used only for telephony and data applications, and are not used for television signal transmission. Satellite channels are also widely used for domestic telephony, data and network television long-haul transmission, as well as television distribution to local outlets. Satellite systems compete favorably with established terrestrial systems for some services. Initially, satellite systems used the 4- and 6-GHz common-carrier bands, sharing frequencies with the terrestrial common-carrier users. The practice has been to use the 6-GHz band to transmit from earth to the satellite and to use the 4-GHz band from the satellite to the earth. In order to have the satellite in a stationary position with respect to points on the earth and be available on a continuous basis, the vehicle must be accurately positioned at a distance of 23,600 miles above

the equator. It was mentioned above that the path loss between antennas increases as the path length increases. At 4 GHz, two isotropic antennas separated 25 miles have a path loss of 137 dB. At 23,600 miles the path loss is 196 dB!

Satellite systems in some respects are quite similar to terrestrial systems. There are radio terminals, transmitters, receivers, and facilities to switch inputs and outputs to and from the various working channels. Specialized power supply equipment that is highly reliable is needed as in terrestrial systems. Because only two ground stations are required to make up a long-haul facility, the financial investment in ground station equipment can be quite high. In enumerating the basic subdivision of satellite system equipment and earth system equipment, the similarities with terrestrial systems end. Specific subunits for satellite systems are quite unique.

In the satellite unit, extreme care in design must be exercised to obtain the ultimate in performance and reliability. Here, two items that are of little or no consequence on earth but are of great importance aloft are weight and power efficiency. Weight is a very important concern in placing the vehicle in orbit, and power efficiency determines how much size, weight, and cost must be dedicated to the solar cell power source. Any trade-off that can be made with the ground station equipment is made, and accordingly, high power in the transmitters, large ground station antennas, and carefully designed high-sensitivity receivers (low-noise figure) are always used. In the past, large klystrons, operating at output levels well over 1000 W, have been used to obtain sufficient power to put an acceptable signal into the satellite receivers. Similarly, large antennas are used to obtain high gain and high directivity of the beam. The latter is important because it is mandatory that a sharp beam be used if other satellites in space are to reuse the limited number of SHF frequencies available. Receiver sensitivity is optimized in the ground station of high-capacity systems usually by refrigerating the front end and achieving the lowest noise figure possible. The importance of the cooling is mentioned in Section VII. M, where it was shown that appreciable improvement in receiver sensitivity (reduction in noise figure) can be achieved using this technique.

Satellites are launched in such a way that, after stabilization in the orbit over the equator, each is assigned a specific longitudinal position that must be maintained. This is called "station keeping" and small jets with onboard gas reservoirs can make minor adjustments of the satellite's position to keep it "on station."

Modern satellites are becoming more sophisticated and include very advanced electronics other than those needed to transmit and receive a SHF signal. Switching of the signals within the satellite permit the routing of circuits on command and provide other operational functions that previously would have been done at ground terminals with much more electronic equipment.

Power in the satellite vehicle is derived from solar energy impinging on the solar cells mounted on the surface of the orbiting station. Some reserve is stored in the batteries to provide power during earth shadows, but the bulk of the power from the solar cells is used directly in the satellite electronics.

Some of the problems that plague terrestrial systems are much less severe or nonexistent in satellite systems. Fading caused by atmospheric conditions along a 25- to 40-mile path in a terrestrial system is virtually nonexistent at 4 and 6 GHz in a satellite path that traverses only about 10 miles of the earth's atmosphere. Therefore, the range of the AGC in the receiver can be much less than in earth systems. Also, the level of the various input signals into the satellite receivers from the earth station transmitters are essentially at equal levels, making the intermodulation problems in the satellite less severe.

The satellite positions (stations) in space are optimized to obtain the maximum number of locations with acceptable levels of interferences from cochannel ground stations. Present practice is to locate satellites at 5° intervals; however, it is proposed that this spacing be reduced to as low as 3° longitude. This results in a finite number of systems that can be deployed using the 4- and 6-GHz bands. Quite a few years ago it was recognized that using only these frequency bands would be inadequate in the years ahead and that systems would have to be developed at higher frequencies, where propagation problems and electronic component design complexities would be more severe. Today, satellite systems operate at 4, 6, 11, and 12 GHz. Some consideration has been given to operating at 25 to 30 GHz.

To achieve the maximum performance of a satellite facility from the standpoint of system capacity and signal-to-noise ratio, the baseband signal is often processed in very specific ways. In the case of television transmission, the electronic elimination of redundancy in the picture allows satisfactory transmission with less baseband occupancy. Compandors (compressors/expandors) are widely used in telephony (but not data) terminals to improve the signal-to-noise ratio in a voice channel by effectively increasing the modulation level on the satellite link. At the present time, frequency modulation of the rf carrier is used in these systems. SSB technology has not yet been employed in satellite systems, but it is now under consideration. The sharing of frequencies with terrestrial systems makes it necessary to coordinate their use in each instance. A terrestrial system normally points the beam very near to the horizon, but if a satellite earth station is in line with this beam, mutual interference can (will) exist. It will be

necessary to relocate one of the sites so that the three stations will not be in this unfavorable alignment. The high power of the earth station transmitter and the high sensitivity of the earth station receiver aggravate the problem in both directions.

XIII. DIGITAL MICROWAVE RADIO

In previous sections, the history of microwave has shown virtually a total use of analog technologies for the transmission of television, telephony, and voiceband data, including telegraph. First, FM was used, and more recently, SSB AM has been introduced primarily to obtain better utilization of the facilities and of the frequency spectrum. To be able to put 6000 voiceband circuits onto a 6-GHz channel when only about 2000 could be accommodated with FM results in a considerable reduction in cost for each circuit carried by the SSB channel. Over the years, it was the primary aim of system designers to increase the load-carrying capacity of the radio channels while still maintaining the quality of the services being rendered.

By contrast, the trend in the local telephone system for the past quarter-century has been to provide interoffice connections of lengths greater than about 5 miles by digital means. This has been an ever-accelerating trend, and the reason for this is the rather large difference in cost of analog multiplex terminals versus digital terminals. What gives the digital terminal the edge over analog techniques is the common use of a large number of components for the digital terminal, whereas an analog multiplexer uses components on a per-circuit basis. Digital facilities are taking over rapidly, and in many instances analog circuits are being removed only to be replaced by digital equivalents. This is done partly to update the equipment in the plant, partly to provide better circuits with lower noise and less maintenance, and partly to be able to handle voice and data services without concern of the mix over the same facility. Once a signal is encoded in digital form, the various kinds of signals are generally indistinguishable from each other and the routing of signals via a digital switch is greatly simplified.

In the mid-1970s, considerable thought was given to whether digital signals could be transmitted economically over microwave channels. It was a well-known fact that a digital voice channel using pulse-code modulation (PCM) requires substantially greater bandwidth than a SSB channel. In the basic PCM terminal, the normal 4-kHz voiceband requires 64 kHz of bandwidth, an increase of 16 times what is required for SSB. It has been stated several times herein that SSB is the most efficient use of the radio spectrum, so the question can be raised, "Why should digital transmission via radio be considered at all?" It has become apparent that the costs of long-haul systems are not only the costs of the radio line but of all of the equipment, including multiplex terminals. More and more switching in the telephone system is being done on a digital basis, and for distances less than about 400 miles, experience shows that digital radio will be less expensive overall than an analog system. (Beyond about 400 miles, the line costs will predominate and it will be more economical to convert digital signals to analog for transmission purposes only.)

Originally, digital modulation schemes were considered wherein phase-shift keying (PSK) of the carrier would convey the information on the channel. Unfortunately, these simple schemes do not provide the bit speeds needed for high system capacity without using excessive bandwidth. A number of other enhancements of the modulation process must be used to obtain higher bit rates. More phases of the carrier can be used, and amplitude shifts plus phase shifts and the modulation of two coherent carriers in QAM can be employed in combination to obtain higher speed of transmission. All of these cause the detection process to become more critical and fragile and require more precise control of all system parameters. Since digital systems are all predicated upon the principle that a pulse will be present or absent at a particular instant, a reference must also be present at the detector so that the received pulse can be recognized when it should be present. Unlike SSB, which requires a precise frequency but not a precise time or phase relationship for proper detection, digital systems must extract precision timing information from the received signal itself. All of this has altered the once-rugged PCM signal to one which is less robust and sensitive to perturbations of the signal than is FM.

With the vast microwave facilities already in place, using well-established channel assignments and channel bandwidths in presently authorized common-carrier bands, digital microwave must fit precisely into the analog microwave world. Furthermore, the FCC has established a rule that a system operating on a 20-MHz-wide channel, as at 4 GHz, shall have the capability of carrying at least 1150 4-kHz voiceband circuits. This requires transmission at a rate of about 78 Mbit/sec or higher. The digital signal must not interfere with adjacent channels that might be operating on other modes of modulation such as FM and SSB AM. Current systems use 90 Mbit/sec for systems with 20-MHz channel bandwidth and $\sim$140 Mbit/sec for systems with 30-MHz channel bandwidths.

It is highly likely that no phase of microwave radio technology has received more attention from the standpoint of high-level mathematical analysis of (1) the frequency spectrum, (2) the error rate performance, and (3) the sensitivity to distortion effects as has digital modulation of

SHF waves. Modern computer programs make possible system simulations and performance evaluations without constructing models and field testing an actual system. This has speeded up the development process and has resulted in more optimum designs.

One of the serious problems encountered in digital radio systems is the dispersive nature of a channel undergoing selective fading. The intersymbol distortion that results when selective fading is present requires the use of space-diversity antennas with combining arrangements or adaptive equalization control of the two paths into the receiver to meet performance objectives. This adds to the complexity of the system and to system costs.

When modulated onto the SHF carrier, a digital signal can produce a frequency spectrum quite different from and at times far broader than either a television signal or a multiplexed telephony signal. The higher the bit rate, the wider the spectrum spread. To keep this under control when operating in a multichannel environment or in a multichannel digital band and to meet FCC spectrum occupancy rules, sophisticated filtering in the IF portions of both the transmitting and receiving equipment of the repeater is used. Again, computer analysis and synthesis help to determine the design of these filters.

XIV. CONCLUSION

After continuous development since the early 1950s of FM analog microwave systems, the law of diminishing returns has set in and further improvements are not likely to be earthshaking. SSB, while much more recent, has already maximized the frequency utilization of 4-kHz occupancy for a 4-kHz voice channel, and this, too, is not likely to show astounding change in capacity. Digital transmission, however, is relatively new, and powerful analytical forces are being brought to bear on all facets of the subject by literally hundreds of researchers and engineers, and it would be highly dangerous to speculate where this fast-moving technology will be in another decade. The future, however, may be made still more uncertain by the rapid growth of fiber guide systems which operate at frequencies very much higher than SHF but which do not experience the anomalies of atmospherically induced dispersion of the transmission medium. This would be a definite advantage for that facility. Permanent contributions have been made since the early 1950s to the quality and the quantity of communications in the United States and throughout the world, and microwave radio has played a large part in bringing about this communications revolution.

SEE ALSO THE FOLLOWING ARTICLES

DATA TRANSMISSION MEDIA • NETWORKS FOR DATA COMMUNICATION NETWORKS • SATELLITE COMMUNICATIONS • RADIO PROPAGATION • RADIO SPECTRUM UTILIZATION

BIBLIOGRAPHY

Carr, J. (1997). "Microwave & Wireless Communications Technology," Newnes, London, UK.

Froehlich, F. E., and Kent, A. (1993). "The Froehlich/Kent Encyclopedia of Telecommunications: Digital Microwave Link Design to Electrical Filters," Volume 6, Marcel Dekker, New York.

Froehlich, F. E., and Kent, A. (1995). "The Froehlich/Kent Encyclopedia of Telecommunications: Microwave Communications Systems and Devices to Modern Optical Character Recognition," Volume 11, Marcel Dekker, New York.

Jakes, W. C. (1994). "Microwave Mobile Communications," IEEE, Edison, NJ.

Larson, L. E. (1996). "RF and Microwave Circuit Design for Wireless Communications," Artech House.

Martin, D. H. (2000). "Communication Satellites," 4th ed., AAAI Press, Menlo Park, CA.

Mehrotra, A. (1994). "Cellular Radio: Analog and Digital Systems," Artech House, Norwood, MA.

Mortazawi, A., and Itoh, T. (1998). "Active Antennas and Quasi-Optical Arrays," IEEE, Edison, NJ.

Pascall, S. C., and Withers, D. (1997). "Commercial Satellite Communications," Focal Press, Woburn, MA.

Saunders, S. R. (1999). "Antennas and Propagation for Wireless Communication Systems," Wiley, New York.

Tsoulos, G. V. (2000). "Adaptive Antennas for Wireless Communications," IEEE, Edison, NJ.

Winch, R. G. (1993). "Telecommunication Transmission Systems: Microwave, Fiber Optic, Mobile Cellular Radio, Data, and Digital Multiplexing," McGraw–Hill, New York.



Modulation

Rodger Ziemer
University of Colorado, Colorado Springs

I. Modulation Basics
II. Analog Modulation
III. Digital Modulation

GLOSSARY

Amplitude-shift keying (ASK) Digital modulation scheme in which information is carried in the amplitude of the carrier.

Analog modulation (AM) Process whereby a message signal, which is the analog of some physical quantity, is impressed on a carrier signal for transmission through a channel.

Analog-to-digital (A/D) conversion The conversion of analog to digital samples.

Antipodal binary signaling Digital modulation scheme in which information is carried in two antipodal states of the transmitted signal.

Bandwidth efficiency The ratio of transmission rate in bits/s to the bandwidth in hertz of a digital modulation scheme.

Carrier signal Signal, usually a sinusoid or a pulse train, which is used to carry the message signal by means of several different types of modulation techniques.

Carson's rule Rule for analog frequency modulation giving the transmission bandwidth in terms of the modulation index and message bandwidth.

Coherent demodulation Recovering a message signal from a modulated signal by making use of a reference signal at the receiver that is in phase coherence with the received carrier.

Combined modulation and coding Any scheme threating modulation and coding together, often to realize simultaneously good performance in terms of power and bandwidth efficiency.

Continuous-phase modulation (CPM) Any digital phase modulation scheme for which the phase changes in a continuous fashion in transitioning from one symbol to the next.

Correlation detector Receiver for digitally modulated signals that correlates a replica of the transmitted signal with the received signal in order to recover the data.

Deemphasis filter Used in conjunction with a preemphasis filter in FM to deemphasize the demodulated message spectrum at higher frequencies to improve interference performance.

Delta modulation Digital pulse modulation scheme that conveys sample values in terms of the differences between successive samples.

Demodulator Device that recovers a message signal from a modulated carrier.

Differential phase-shift keying (DPSK) Digital modulation scheme that conveys information symbols as phase differences of the carrier.

Digital modulation Technique for placing a digital data sequence on a carrier signal for subsequent transmission through a channel.

Direct sequence spreading Spread spectrum modulation

scheme that uses pseudorandom sequences to spread the spectrum through phase modulation.

Envelope detector Device producing an output proportional to an AM signal envelope.

Frequency hop spreading Spread spectrum modulation scheme that uses pseudorandom frequency hopping to spread the spectrum through frequency modulation.

Frequency-shift keying (FSK) Digital modulation scheme that conveys information symbols as frequency shifts of the carrier. Can be either coherent or noncoherently detected.

Gaussian minimum-shift keying (GMSK) MSK system using Gaussian pulse shaping for the phase changes.

Intersymbol interference (ISI) Interference arising from spreading of one symbol into the next in digital modulation, often due to unwanted filtering by the channel.

Linear modulation Modulation technique in which the modulated carrier is a linear function of the message signal.

Lower sideband Portion of a modulated signal spectrum with frequency components having frequencies less than the carrier frequency.

Maximum likelihood detection Detection scheme that decides in favor of the most likely symbol to have been transmitted.

Message signal Signal that contains the information to be transmitted through the channel through modulation of a carrier signal.

Minimum-shift keying (MSK) Digital modulation scheme where information is conveyed by $\pm\pi/2$ radian linearly varying phase changes of the carrier over the symbol interval.

Modulation Process whereby a message signal is impressed on a carrier signal for subsequent transmission through a channel.

Multiplexing Modulation process that makes it possible to transmit several message signals through the same channel on a single carrier.

Noncoherent modulation Digital modulation scheme where the digital data is recoverable from the modulated signal without the use of phase-coherent reference signal at the receiver.

Nonlinear analog modulation Modulation scheme for which the modulated carrier is a nonlinear function of the message signal.

Nyquist's pulse-shaping criterion Gives the condition for a strictly bandlimited pulse spectrum that corresponds to a pulse shape with zeros at symbol-spaced intervals (zero ISI).

Nyquist's sampling theorem It states that the minimum sampling rate (samples per second) of an analog lowpass signal must be at least twice the signal's bandwidth in hertz in order for it to be recoverable from the sample values.

Offset quadrature phase-shift keying (OQPSK) QPSK modulation scheme with quadrature data streams offset relative to each other by a half symbol interval.

Optical communications Any communication system where the information is conveyed by light waves.

Phase-lock loop (PLL) Device that provides a coherent phase reference through a feedback structure involving a phase detector and voltage-controlled oscillator.

Phase-shift keying (PSK) Digital modulation scheme wherein information is conveyed through phase shifts of the carrier.

Power efficiency The signal-to-noise ratio required for a digital modulation scheme to provide a prespecified probability of error.

Preemphasis filter Used with a deemphasis filter in FM to boost the message spectrum at higher frequencies before modulation to improve interference performance.

Probability of error Performance measure in a digital modulation scheme—the ratio of number of symbols received in error to the total number received over a time period that is sufficiently long that the ratio is essentially constant.

Pulse amplitude modulation (PAM) Analog pulse modulation scheme where information is conveyed in the pulse amplitudes of a pulse-train carrier.

Pulse-code modulation (PCM) Modulation scheme where message sample values are conveyed as a binary code after A/D conversion.

Pulse position modulation (PPM) Analog pulse modulation scheme where information is conveyed in terms of pulse positions relative to the start of the sampling interval.

Pulse width modulation (PWM) Analog pulse modulation scheme where information is conveyed in terms of pulse widths of a pulse-train carrier.

Quadrature-amplitude modulation (QAM) A digital modulation scheme wherein symbols are conveyed as combinations of pulse amplitude values of the carrier.

Quadrature phase-shift keying (QPSK) A digital modulation scheme wherein a sine carrier is modulated by one antipodal symbol stream, a cosine is modulated by another, and the two are summed. Symbol streams may be from separate sources or originate from one source.

Raised cosine spectrum A signal spectrum that satisfies Nyquist's pulse shaping criterion.

Shannon's capacity formula Gives the capacity of an AWGN channel in terms of signal-to-noise ratio; at rates below capacity, error-free performance is theoretically possible.

Sidebands Paired spectral components carry duplicate information symmetrically placed about the carrier frequency of the modulated carrier.

Single-sideband (SSB) modulation Analog modulation scheme wherein only one sideband is transmitted.

Slope detector FM detector implemented by a circuit with linear slope about the carrier frequency, thereby converting frequency excursions from the carrier to envelope variations.

Spectrum Function that gives the relative magnitudes and phases of a signal (called phasor components) versus frequency such that the signal can be reconstructed through vector addition of these phasor components.

Square-root-raised cosine Refers to the ideal frequency response of the spectral shaping filters at transmitter and receiver to be implemented in order to allow strictly bandlimited transmission in a digital system without introducing intersymbol interference.

Turbo coding A recently discovered coding scheme that gets very close to the Shannon capacity bound.

Upper sideband Portion of the spectrum of a modulated signal whose frequencies are greater than the carrier frequency.

Vestigial-sideband (VSB) modulation Essentially single-sideband with a vestige of the eliminated sideband remaining.

Voltage-controlled oscillator (VCO) An oscillator circuit producing an output, often sinusoidal but that can also be a pulse signal, whose frequency is proportional to an input voltage.

MODULATION is the systematic variation of some attribute of a carrier signal in accordance with a function of the message signal. Examples of carrier signals are sinusoids or pulse trains. Various attributes of sinusoids that can be varied in accordance with the message are amplitude, phase, or frequency, while those of a pulse train are pulse amplitude, pulse width, or pulse position with respect to a fixed reference. Message signals can be the analog of some quantity, such as a speech wave, in which case the modulation is referred to as analog. If the message is chosen from a discrete set, such as the English alphabet, and a carrier attribute, such as phase, is varied in a one-to-one manner in accordance with this discrete message, the modulation is referred to as digital. Modulation techniques can be further categorized as linear or nonlinear, depending on whether the modulated signal varies linearly with the message (i.e., superposition holds) or whether it varies nonlinearly with the message. Examples of the former are amplitude modulation, a term usually applied to analog modulation, or amplitude-shift keying, which is the term used when the modulation is digital. Examples of nonlinear modulation schemes include phase and frequency modulation, which are usually the terms applied to nonlinear analog modulation, or phase-shift keying and frequency-shift keying in the case of digital modulation. Reasons for modulation include (1) easing radiation of the signal by means of an antenna; (2) reducing noise and interference; (3) assigning channels; (4) multiplexing in order to transmit many messages through the same channel at the same time; and (5) overcoming equipment limitations. The process of modulation is carried out by means of a modulator and the inverse operation of recovering the signal from the modulated signal is carried out by a demodulator.

I. MODULATION BASICS

A. Signals and Spectra

The study of modulation methods is facilitated by using both time and frequency domain representations. A signal's frequency domain description, or spectrum, is given by

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft)\, dt \triangleq \mathcal{F}[x(t)] \qquad (1)$$

The time domain representation of the signal $x(t)$ is obtained from its spectrum by the integral

$$x(t) = \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft)\, df \triangleq \mathcal{F}^{-1}[x(t)] \qquad (2)$$

which is called the inverse Fourier transform of $X(f)$. These integrals provide what is referred to as a Fourier transform pair of any signal. Not only can tables of Fourier transform pairs be derived, but theorems relating important properties of Fourier transforms can be obtained and give much insight into how various operations on signals affect their spectra. Important theorems for the consideration of modulation techniques are the frequency translation and modulation theorems, which state the following. Given a signal $x(t)$ with Fourier transform $X(f)$, then

$$\mathcal{F}[x(t) \exp(\pm j2\pi f_0 t)] = X(f \pm f_0)$$

frequency translation theorem (3)

and

$$\mathcal{F}[x(t) \cos(2\pi f_0 t)] = \frac{1}{2}[X(f - f_0) + X(f + f_0)]$$

modulation theorem (4)

These theorems simply state that multiplication of a signal by a sinusoid of constant frequency f_0 shifts the spectrum of the signal from being centered around zero frequency to being centered around frequency f_0.

Convenient Fourier transform pairs for the sequel are

$$\mathcal{F}[A\cos(2\pi f_0 t)] = \frac{A}{2}[\delta(f - f_0) + \delta(f + f_0)] \quad (5)$$

and

$$\mathcal{F}[A] = A\delta(f) \quad (6)$$

where $\delta(f)$ is the unit impulse or delta function and A is a constant. These transform pairs, as well as the theorems given previously, will prove useful in the description of modulation systems.

B. Analog Modulation Terminology and Methods

In analog modulation systems, the transmitted carrier is modulated by a signal that is the analog of some physical quantity, such as a speech wave or velocity. Analog message signals are often used to modulate sinusoidal carriers of the form

$$x_c(t) = A_c\cos(2\pi f_c t + \theta_c) \quad (7)$$

Chief means of analog modulation of sinusoidal carriers include double-sideband modulation (DSB), amplitude modulation (AM), frequency modulation (FM), and phase modulation (PM). In the first two cases, the amplitude of the carrier is varied linearly in accordance with the message signal. In the case of FM, the instantaneous frequency of the carrier is varied linearly as a function of the message signal while, for PM, the instantaneous phase of the carrier is varied in proportion to the message signal.

Continuous-time signals are sometimes sampled to produce discrete-time signals, also referred to as sample-data signals. Sample-data message signals are often used to modulate pulse train carriers of the form

$$p_c(t) = A_c\sum_{n=-\infty}^{\infty}\Pi[(t - nT_s - t_0)/\tau] \quad (8)$$

where $\Pi(t) = 1$ for $|t| \le 1/2$ and 0 otherwise. If the amplitude, pulse position, or pulse width is varied in proportion to the message sample values, the result is pulse amplitude modulation (PAM), pulse position modulation (PPM), or pulse width modulation (PWM), respectively.

C. Digital Modulation Terminology and Methods

Digital modulation results if a new message is chosen each T_s seconds (called the symbol interval) from a discrete set and mapped one-to-one into a signal parameter value (amplitude, phase, etc.). If only two messages are possible, the modulation is called binary. If $M > 2$ messages are possible, it is referred to as M-ary. It is convenient to denote the message set as {0, 1} for binary whereas, for M-ary modulation, it is convenient to denote it as $\{0, 1, \ldots, M-1\}$.

If the carrier is sinusoidal, the amplitude, phase, or frequency can be varied in accordance with the message set, in which case the modulation is referred to as amplitude-shift keying (ASK), phase-shift keying (PSK), or frequency-shift keying (FSK), respectively. Other higher-order digital modulation schemes can also be employed. For example, if both amplitude and phase are varied, the result is referred to as quadrature-amplitude modulation (QAM). In phase modulation, a continuous phase change from one symbol interval to the next results in continuous-phase modulation (CPM). A special case of binary CPM results when the phase changes linearly with a total phase change in a symbol interval being $\pi/2$ radians, which results in minimum-shift keying (MSK). More complex CPM modulation schemes can be used to provide the simultaneous desirable attributes of constant amplitude, bandwidth efficiency, and power efficiency (for example, a nonlinear phase variation over several symbol intervals).

With the advent of laser sources and low-loss optical fibers, digital communications in the optical portion of the spectrum has become the mode of choice for land-line voice and data communications. Initially, digital optical communications made use of photodetectors, wherein a burst of light energy is directly detected and output from the detector is a baseband voltage pulse. More recent forms of optical communications make use of coherent modulation and wavelength-division multiplexing, analogous to frequency-division multiplexing, which allows data rates of the order of several gigabits per second.

II. ANALOG MODULATION

A. Linear Analog Modulation Schemes

1. Double-Sideband and Amplitude Modulation

In linear analog modulation, the amplitude of the carrier signal depends linearly on the analog message signal. Three examples of linear modulation are DSB, AM, and PAM. PAM is discussed later with the other possible analog pulse modulation methods. In the case of AM and DSB, the carrier is sinusoidal as expressed by (7) and the modulated signal is written as

$$x_{c,\mathrm{DSB}}(t) = A_c m(t)\cos(2\pi f_c t) \quad (9)$$

for DSB, and

$$x_{c,\mathrm{AM}}(t) = A_c[1 + am(t)]\cos(2\pi f_c t) \quad (10)$$

for AM, where $m(t)$ is the message signal, f_c is the carrier frequency, and a is a parameter called the modulation

index. A modulator for either DSB or AM is implemented as a multiplier.

If the Fourier transforms of Eqs. (9) and (10) are taken, the resulting spectra are

$$X_{c,DSB}(f) = \frac{1}{2}A_c[M(f - f_c) + M(f + f_c)] \quad (11)$$

for DSB, and

$$X_{c,AM}(f) = \frac{1}{2}A_c[\delta(f - f_c) + \delta(f + f_c)] + \frac{1}{2}aA_c[M(f - f_c) + M(f + f_c)] \quad (12)$$

for AM. The main difference between DSB and AM is the presence of a carrier component in AM, which manifests itself as the term $A_c \cos(2\pi f_c t)$ in the time domain and as the pair of delta functions located at frequencies $\pm f_c$ in the frequency domain. Waveforms and spectra for DSB and AM modulation where the message signal is sinusoidal are compared in Fig. 1. Note that in both cases, the spectrum corresponding to the message signal possesses a mirror-image symmetry about the carrier (the carrier is absent in the case of DSB). These paired message spectral components are known as the upper and lower sidebands and carry duplicate information.

The demodulation of either DSB or AM can be accomplished by a coherent demodulator shown schematically in Fig. 2a. The implementation of such a demodulator presupposes the availability of a coherent carrier reference at the receiver. A circuit for achieving this is the squarer-frequency-divider circuit of Fig. 2b; the Costas loop circuit of Fig. 2c establishes a coherent reference and provides the demodulated signal at its output.

The use of coherent demodulation is necessary for the successful demodulation of DSB because of the 180°

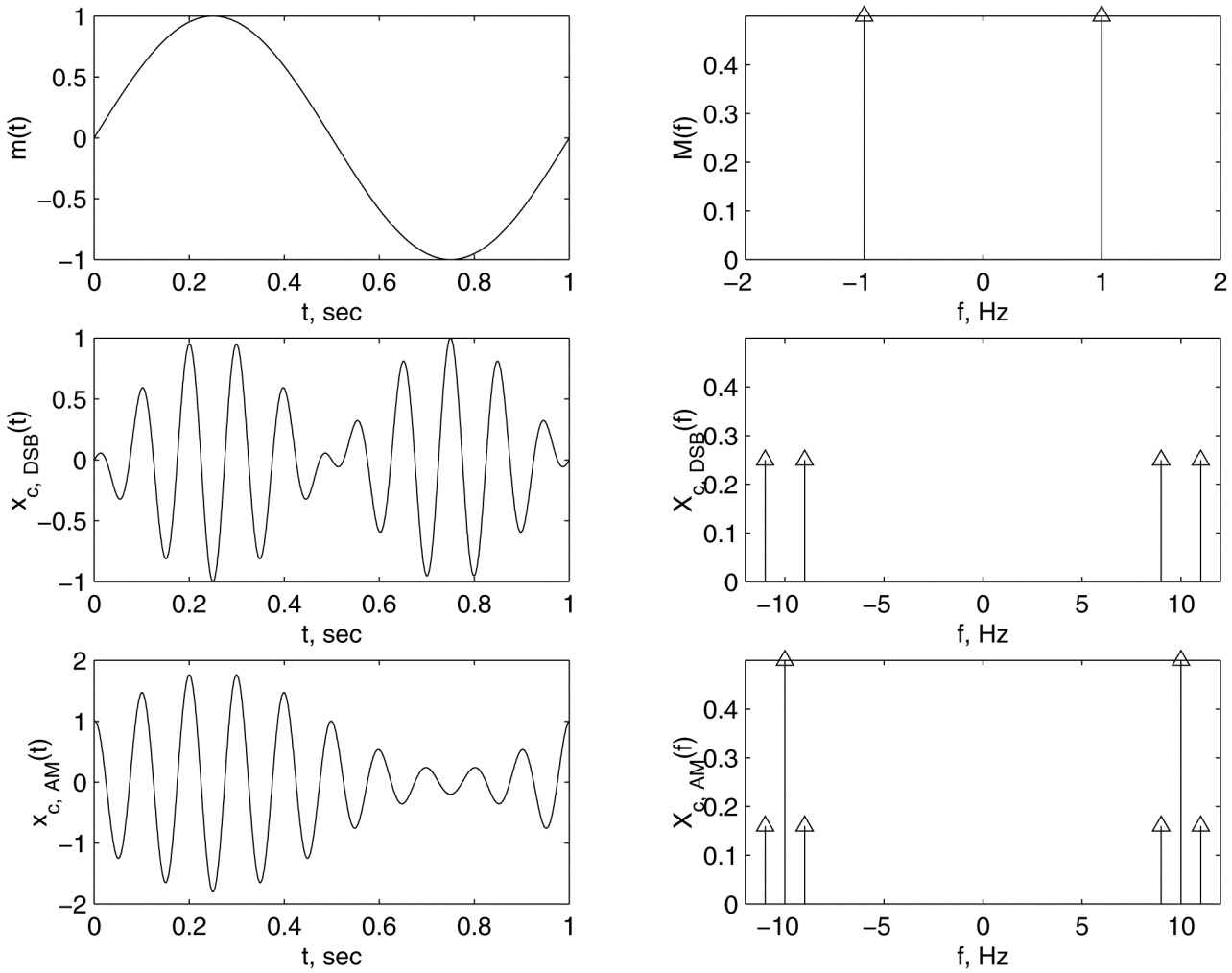


FIGURE 1 Waveforms and spectra for (a) DSB modulation and (b) AM modulation ($a = 0.8$); sinusoidal modulating signal frequency of 1 Hz and carrier frequency of 10 Hz.

phase reversals of the carrier that occur each time $m(t)$ changes sign. Because of the presence of a carrier component in AM, these phase reversals do not occur if $a > 1$, thus allowing the use an envelope detector as illustrated in Fig. 2d. The diode passes the positive halves of the modulated carrier sine wave, and these pulses are then smoothed into a replica of the modulating signal by the low-pass RC filter that follows the rectifer. The simplicity of this scheme is one reason for its use in commercial AM broadcast radio.

2. Single-Sideband and Vestigial-Sideband Modulation

Two modulation schemes that are derivatives of DSB modulation are single-sideband (SSB) and vestigial-sideband (VSB) modulation. SSB modulation is produced by filtering out all of one sideband of a DSB signal, while VSB is produced by leaving a vestige of one sideband and all of the other sideband of a DSB signal. Such sideband rejection can be effected by a sideband filter at the output of the DSB modulator, which consists of a multiplier as pointed out previously. Clearly, SSB conveys the information in the message signal while halving the bandwidth compared with DSB. The reason for leaving a part of one sideband in VSB is to allow for easier sideband filter realization. Both SSB and VSB can be demodulated using the coherent demodulator shown in Fig. 2a. Both modulation schemes may also be demodulated by reinserting a carrier component and passing the sum through an envelope detector. VSB is used for transmitting the video signal in commercial television broadcasting.

B. Nonlinear Analog Modulation Schemes

Examples of nonlinear analog modulation schemes are FM, PM, PWM, and PPM. PWM and PPM will be discussed later. For the sinusoidal carrier of (7), the instantaneous phase is

$$2\pi f_{\mathrm{i}}(t) = \frac{d\theta_{\mathrm{i}}(t)}{dt} = \frac{d}{dt}[2\pi f_{\mathrm{c}} t + \theta_{\mathrm{c}}(t)] = 2\pi f_{\mathrm{c}} + \frac{d\theta_{\mathrm{c}}(t)}{dt} \tag{13}$$

where $\theta_{\mathrm{c}}(t)$ and $d\theta_{\mathrm{c}}(t)/dt$ are the phase and frequency deviations, respectively. If the phase deviation is proportional to the message signal the result is PM, whereas if the frequency deviation is proportional to the message signal, FM results. Thus, for PM, the modulated signal is

$$x_{\mathrm{c,PM}}(t) = A_{\mathrm{c}} \cos[2\pi f_{\mathrm{c}} t + k_{\mathrm{p}} m(t)] \tag{14}$$

whereas for frequency modulation, the resultant modulated signal is of the form

$$x_{\mathrm{c,FM}}(t) = A_{\mathrm{c}} \cos\left[2\pi f_{\mathrm{c}} t + k_{\mathrm{f}} \int^{t} m(\alpha)\, d\alpha\right] \tag{15}$$

where k_{p} and k_{f} are proportionality constants. Both FM and PM are referred to as angle-modulated signals.

Because of the nonlinear dependence of the modulated signal on the message signal for FM and PM, general results for their spectra are difficult to obtain. A specific case is for a sinusoidal message signal of amplitude A_{m} and frequency f_{m} for which (14) and (15) take the form

$$x_{\mathrm{c}}(t) = A_{\mathrm{c}} \cos[2\pi f_{\mathrm{c}} t + \beta \sin(2\pi f_{\mathrm{m}} t)] \tag{16}$$

in which β is referred to as the modulation index, given by

$$\beta = \frac{A_{\mathrm{m}} f_{\mathrm{m}}}{f_{\mathrm{d}}} \text{ for FM where } f_{\mathrm{d}} = k_{\mathrm{f}}/2\pi \tag{17}$$

and

$$\beta = k_{\mathrm{p}} A_{\mathrm{m}} \text{ for PM} \tag{18}$$

Equation (16) can be expanded in the trigonometric series

$$x_{\mathrm{c}}(t) = \sum_{n=-\infty}^{\infty} A_{\mathrm{c}} J_{\mathrm{n}}(\beta) \cos[2\pi(f_{\mathrm{c}} + n f_{\mathrm{m}})\, t] \tag{19}$$

where $J_{\mathrm{n}}(\beta)$ is a Bessel function of order n and argument β. Because of the doubly infinite series, it is evident that for sinusoidal modulation the spectrum of an angle-modulated signal is infinite in extent and consists of lines at the frequencies f_{c}, $f_{\mathrm{c}} \pm f_{\mathrm{m}}$, $f_{\mathrm{c}} \pm 2f_{\mathrm{m}}$, . . . , etc.

Figure 3 shows amplitude spectra for a sinusoidally angle-modulated carrier for various values of the modulation index β. Note that the spectrum is negligible beyond a certain frequency separation from the carrier. A rule of thumb often used to estimate the transmission bandwidth, B_{T}, required for an angle modulated signal is Carson's rule, which is

$$B_{\mathrm{T}} = 2(\beta + 1) f_{\mathrm{m}} \tag{20}$$

This rule is heuristically applied to the case of nonsinusoidal modulating signals by replacing β with a deviation ratio, D, defined as

$$D = \frac{\text{peak frequency deviation}}{\text{bandwidth of } m(t)} = \frac{f_{\mathrm{d}}}{W}[\max|m(t)|] \tag{21}$$

where W is the bandwidth of $m(t)$; in (20), f_{m} is also replaced by W.

The demodulation of FM can be accomplished by a *slope detector*, which, in its simplest form, can be modeled as a differentiator followed by an envelope detector. That such a circuit can be used for demodulation of FM can be seen by differentiating (15), which gives

$$y(t) = -A_{\mathrm{c}}[2\pi f_{\mathrm{c}} + k_f m(t)] \sin\left[2\pi f_{\mathrm{c}} t + k_{\mathrm{f}} \int^{t} m(\alpha)\, d\alpha\right] \tag{22}$$

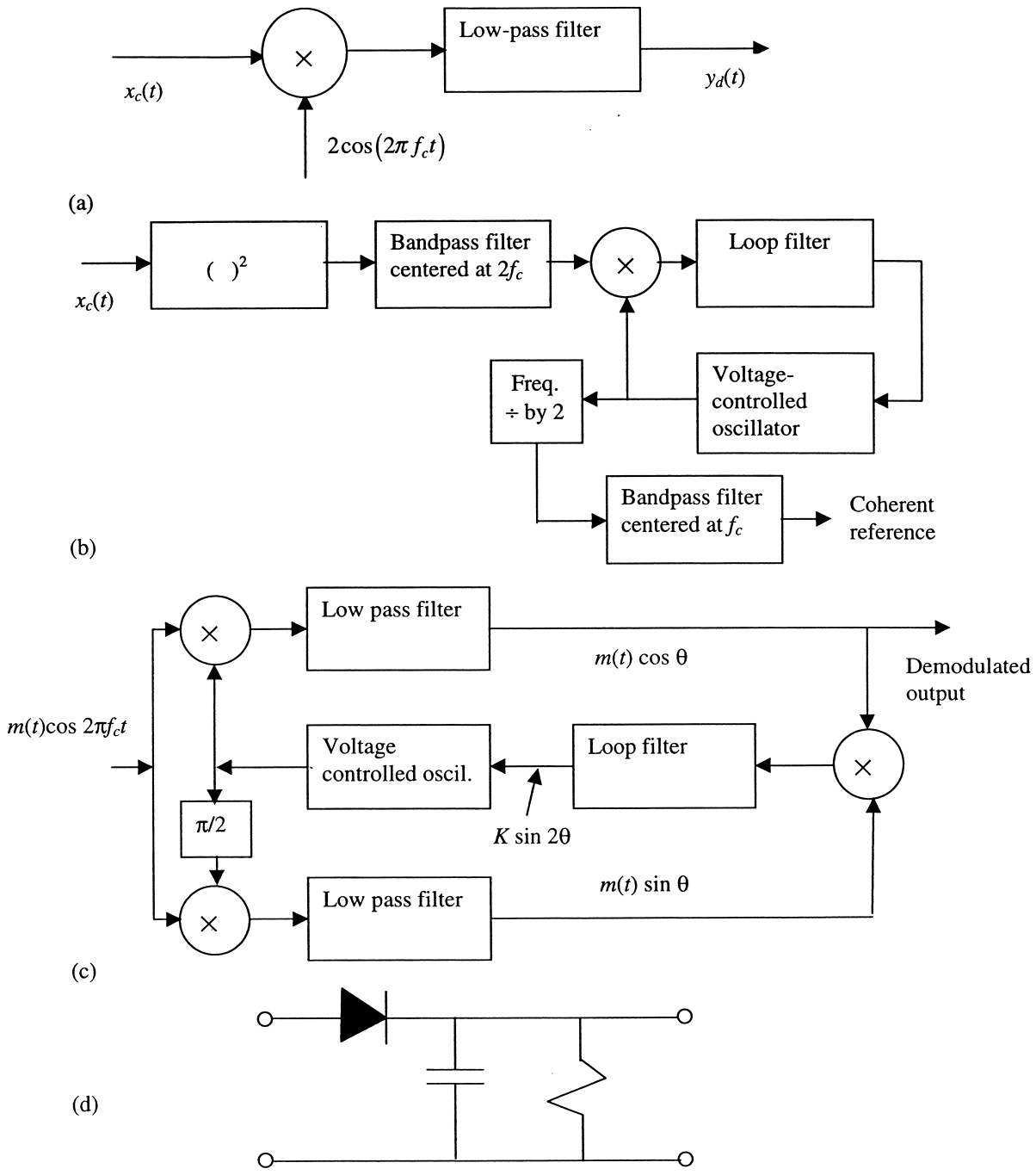


FIGURE 2 (a) Coherent demodulator for DSB and AM; (b) squarer-frequency-divider to obtain reference for demodulating DSB; (c) Costas loop to demodulate DSB; (d) envelope detector.

The envelope of (22) is

$$e(t) = A_c[2\pi f_c + k_f m(t)] \tag{23}$$

which is proportional to the message, $m(t)$. In actuality, any circuit with a frequency response that is approximately linear (nonzero slope) in a frequency range around f_c will do in place of the differentiator. To demodulate PM, one can follow the envelope detector with an integrator circuit. Nowadays, a phase-lock loop (PLL) circuit is typically used to demodulate FM because of the availability of inexpensive PLL integrated circuits.

Frequency modulation is used for commercial broadcast radio with a carrier frequency in the range of 88.1

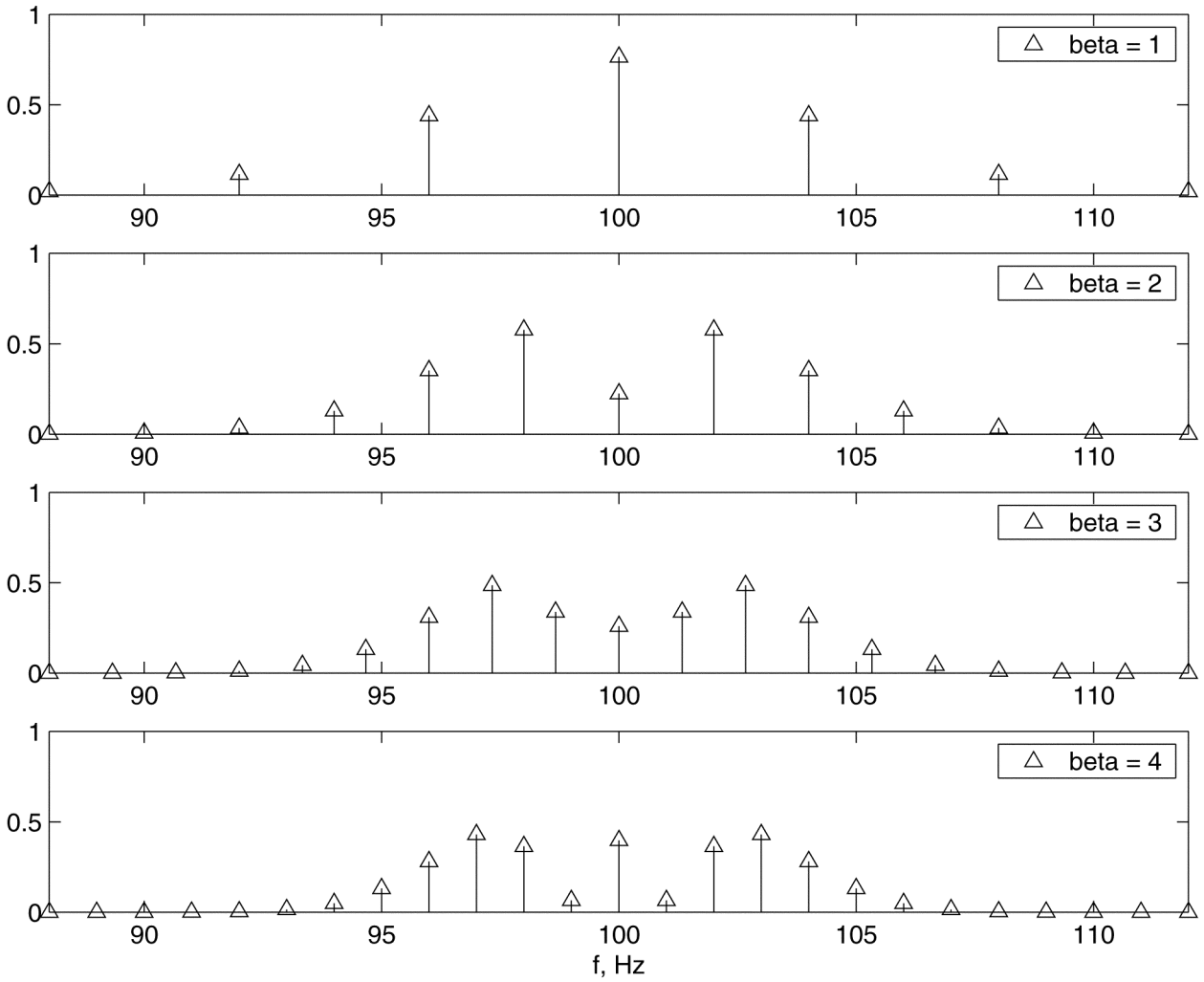


FIGURE 3 Amplitude spectra for a sinusoidally modulated FM signal ($\beta f_m = 4$ Hz).

to 107.9 MHz and a channel spacing of 200 kHz. In this application, a technique known as pre- and deemphasis is used to combat the effects of interference, which is most detrimental in FM at the higher frequency end of the message spectrum. At the transmitter, a preemphasis filter is used to boost the high-frequency end of the message spectrum. At the receiver, after demodulation, a deemphasis filter is used to bring the high-frequency end of the message spectrum down to what it was at the transmitter while simultaneously attenuating noise at frequencies at the high-frequency end of the message spectrum where they are most severe.

C. Analog Pulse Modulation

Recall that the three types of pulse-train carrier modulation, with carrier of the form (8), are PAM, PWM, and PPM. In PAM, the amplitude of the pulse carrier is modulated by the samples of the message signal taken at T_s-second intervals at a rate $f_s = 1/T_s \geq 2W$ samples/second according to Nyquist's sampling theorem, where W is the bandwidth of the $m(t)$. The PAM-modulated waveform is

$$x_{c,\mathrm{PAM}}(t) = A_c \sum_{n=-\infty}^{\infty} m(nT_s)\Pi[(t - nT_s)/\tau] \qquad (24)$$

where the pulse width τ is held constant. The PAM signal (24) is called flat-top sampling, and the message can be recovered by a low-pass filter of bandwidth greater than W but less than $f_s - W$. Ideally, the low-pass filter should be followed by an equalizer, which removes the distortion introduced by the flat-top pulse samples. In this idealized case, the equalizer has a frequency response that is the

inverse of the spectrum of the rectangular pulse, or the inverse of

$$\tau \text{sinc}(f\tau) \equiv \tau \frac{\sin(\pi f \tau)}{\pi f \tau}.$$

Equalization is unnecessary if $\tau \ll T_s$.

The other two types of analog pulse modulation, PWM and PPM, are nonlinear modulation techniques as contrasted to PAM, which is linear. As such, their baseband spectra are significantly wider than that of $m(t)$. This disadvantage is counterbalanced by the advantage that they are more immune to the detrimental effects of noise than PAM. PWM can be demodulated by low-pass filtering, just as for PAM. PPM can also be demodulated by low-pass filtering after converting it to PWM with the aid of a reference clock signal.

D. Pulse Code and Delta Modulation

Representation of analog messages in digital format is effected by first sampling the message signal, $m(t)$, in accordance with the sampling theorem and then converting the samples to binary number representation through the process of analog-to-digital (A/D) conversion. For example, for the message samples given in Table I, the corresponding A/D converted representation is given in the last column (rounding to the nearest integer value). The binary representation of each sample value is then transmitted through the channel by a sequence of on–off rectangular pulses in the T_s-second interval corresponding to that sample value. This procedure is referred to as pulse-code modulation (PCM).

Another way of representing the message sample values in binary form is by means of delta modulation. This technique uses a feedback modulator structure that compares a binary representation of the current sample value with the actual value. If greater than the actual value, a logic-1 is sent through the channel; if less than the actual value, a logic-0 is sent through the channel. This is the simplest version of delta modulation, known as zero order. Higher-order versions take also use the slope and/or curvature of the message signal at the sampling instant.

TABLE I Example of PCM Waveform Representation

Sample time	Sample value	Rounded value	Binary representation
T_s	1.1	1	001
$2T_s$	2.4	2	010
$3T_s$	3.1	3	011
$4T_s$	1.9	2	010
$5T_s$	5.1	5	101
$6T_s$	6.6	7	111
$7T_s$	7.4	7	111

III. DIGITAL MODULATION

A. Performance Measures

Two primary considerations when choosing a digital modulation scheme are bandwidth occupancy, measured in terms of information rate per unit bandwidth (bits/sec/Hz), and error-rate performance, measured in terms of the signal energy-to-noise-spectral-density ratio (E_b/N_0), required to give a specified probability of bit error, P_b. The first presupposes that some means of specifying bandwidth has been chosen. This can be in terms of an energy containment bandwidth, which is the bandwidth within which a specified percent of the total energy is contained, or the bandwidth required for the main lobe of the signal spectrum. The former is more precise, and the latter easier to apply. For this reason, the latter will be used in the following discussion even though the numbers given are somewhat pessimistic.

The trade-off between power and bandwidth efficiency is fundamental, being governed in the ideal case of additive white Gaussian noise (AWGN) by Shannon's capacity formula, which is

$$\frac{C}{W} = \log_2\left(1 + \frac{P}{N_0 W}\right) = \log_2\left(1 + \frac{E_b}{N_0}\frac{R}{W}\right) \text{ bits} \tag{25}$$

where C is the channel capacity (bits/s), W is the transmission bandwidth (Hz), $P = E_b R$ is the signal power (W), N_0 is the single-sided noise power spectral density (W/Hz), E_b is the energy per bit of the received signal (J), and R is the data rate (bits/s).

If $R < C$, Shannon showed that error-free transmission through the channel is possible, although at the cost of complex coding and long delay. No signaling scheme can convey information from source to receptor without inevitable errors for $R > C$. Practical digital modulation schemes to be discussed shortly fall far short of achieving $R = C$, although a coding technique called turbo coding comes much closer than any scheme known up the time of its discovery in 1993 (Berrou, Glavieux, and Thitimajshima, 1993).

Secondary criteria for choosing modulation schemes are ease of implementation and constant envelope of the modulated signal, which is quite often of importance if limiting amplifiers are used for efficiency reasons.

B. Binary versus *M*-ary Schemes

One trade-off available to the digital communication system designer is the number of possible signals selected

from each signaling interval. An M-ary system conveys $\log_2 M$ bits per signaling interval. For example, an 8-ary system selects one of eight possible symbols each signaling interval which may be associated with one of the following bit sequences: 000, 001, 010, 011, 100, 101, 110, and 111. The choice of M is decided on the basis of power and bandwidth efficiency trade-off, complexity of implementation, constant versus nonconstant envelope, etc. The question of power and bandwidth efficiency can most fairly be decided on the basis of required E_b/N_0 for a desired equivalent bit error probability and data rate in terms of bits/s/Hz, respectively.

The trade-off between power and bandwidth efficiency is intuitively visualized in terms of signal space. It can be shown that an orthogonal basis function set can be found for any M-ary modulation scheme, allowing the possible transmitted signals to be represented geometrically as

$$s_i(t) = \sum_{k=1}^{K} S_{ik}\phi_k(t)$$
$$i = 1, 2, \ldots, M; \quad K \le M; \quad 0 \le t \le T_s \qquad (26)$$

where the orthogonal basis set $\{\phi_k(t), k = 1, 2, \ldots, K\}$ defines the coordinate system and $\{S_{ik}\}$ are the coordinates of the ith signal. By orthogonal, it is meant that

$$\int_0^{T_s} \phi_i(t)\phi_i(t)\,dt = \begin{cases} 1, i = j \\ 0, i \ne j \end{cases} \qquad (27)$$

It is assumed that the demodulators/detectors are optimum in the sense that they achieve the minimum error probability in additive AWGN of all possible detectors. They are generally known as maximum likelihood detectors and involve a bank of matched filters or correlators followed by threshold devices and comparison logic. In AWGN, it can be shown that this type of detector computes the signal space coordinates of the received signal plus noise and chooses as the most likely transmitted signal the one closest in Euclidean distance to the received signal plus noise. Figure 4 shows a block diagram for accomplishing this optimum detection strategy.

C. Coherent Digital Modulation Schemes

A coherent digital modulation scheme requires a reference signal at the receiver that is in phase coherence with the carrier of the transmitted signal (which may be suppressed), accounting for the phase shifts introduced by the channel. Noncoherent schemes require no such reference signal, but at the expense of poorer power efficiency. The price paid for the improved performance of coherent schemes is additional complexity of the reference acquisition circuit.

The characteristics of several coherent digital modulation schemes are summarized in Table II. In this table, the following definitions apply: E_b is the symbol energy for a binary scheme; E_s is the symbol energy for an M-ary

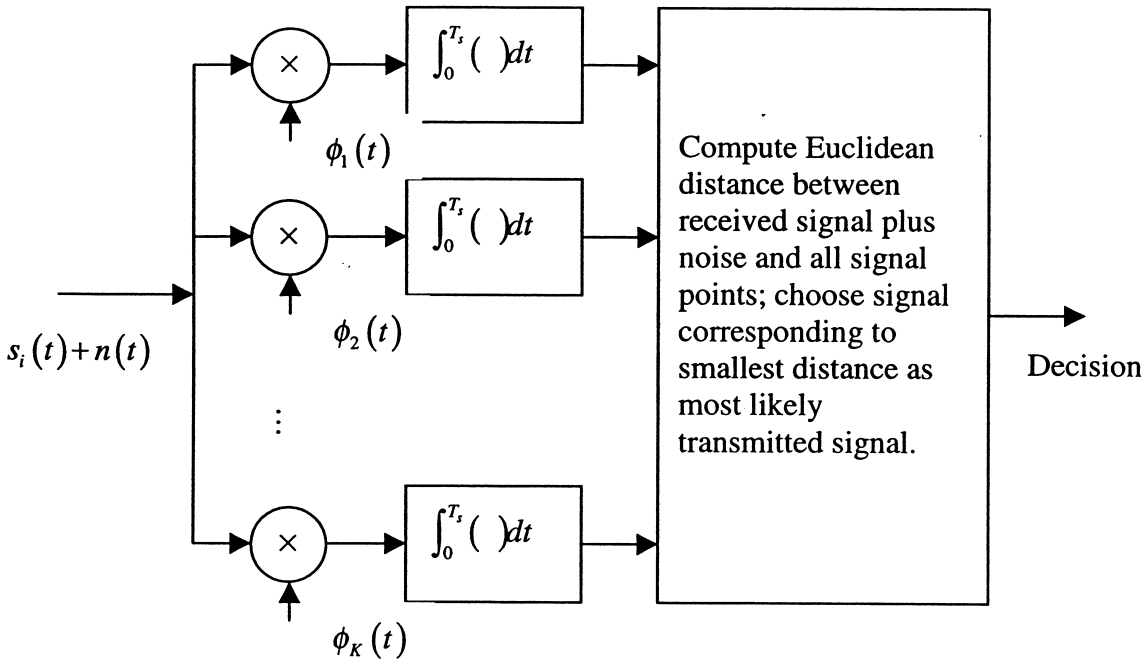


FIGURE 4 Correlation detector that implements the maximum likelihood detection criterion (coherent signaling).

TABLE II Characteristics of Various Digital Modulation Schemes

Modulation type	Signal set [signaling interval: $(i-1)T_s \le t \le iT_s$]	Sig. Sp. dim.	BW eff. (bps/Hz)	Probability of bit error, P_b	E_b/N_0 (dB) for $P_b = 10^{-6}$
Binary ASK	$s_i(t) = 2(i-1)\,a\cos(\omega_c t + \theta)$; θ uniform in $[0, 2\pi]$ $i = 1, 2, \ldots, M$; $a^2 = \frac{3\bar{E}_s}{2(M-1)(2M-1)}$	1	0.5	$\frac{2(M-1)}{M\log_2 M}\,Q\left(\sqrt{\frac{3\log_2 M}{(M-1)(2M-1)}\,\frac{\bar{E}_b}{N_0}}\right)$	$M=2$; 13.55 $M=4$; 26 $M=8$; 38.1
QPSK; OQPSK	$\sqrt{E_s/T_s}[d_1(t)\cos\omega_c t + d_2(t)\sin\omega_c t]$ d_1 and d_2 offset by $T_b = T_s/2$ for OQPSK	2	1.0	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$	10.54
MSK	$\sqrt{\frac{2E_s}{T_s}}\left[d_1(t)\cos\frac{\pi t}{T_s}\cos\omega_c t + d_2(t)\sin\frac{\pi t}{T_s}\sin\omega_s t\right]$	2	0.67	$Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$	10.54
GMSK	$x_c(t) = \sqrt{2P_c}\cos\left[2\pi f_0 t + {}_2\pi f_d \sum_{n=-\infty}^{\infty} a_n g(t-nT_b)\right]$, $\{a_n\} =$ bits $g(t) = \frac{1}{2}\left[\mathrm{erf}\left(-\sqrt{\frac{2}{\ln 2}}\pi BT_b\left(\frac{t}{T_b}-\frac{1}{2}\right)\right) + \mathrm{erf}\left(\sqrt{\frac{2}{\ln 2}}\pi BT_b\left(\frac{t}{T_b}+\frac{1}{2}\right)\right)\right]$	2	Depends on BT_b	See text commentary	Depends BT_b See text
M-ary PSK	$s_i(t) = \sqrt{\frac{2E_s}{T_s}}\cos\left[\omega_c t + \frac{2\pi(i-1)}{M}\right]$; $i = 1, 2, \ldots, M$	2	$\frac{\log_2 M}{2}$	$< \frac{2}{\log_2 M}\,Q\left(\sqrt{\frac{2(\log_2 M)E_b}{N_0}}\right)$	$M=2, 4$; 10.5 $M=8$; 13.7 $M=16$; 18.2
Coherent M-ary FSK	$s_i(t) = \sqrt{\frac{2E_s}{T_s}}\cos[\omega_c + (i-1)\Delta\omega]t$	M	$\frac{\log_2 M}{M+1}$	$\approx \frac{M}{2}\,Q\left(\sqrt{\frac{2(\log_2 M)E_b}{N_0}}\right)$ exact for $M=2$	$M=2$; 13.6 $M=4$; 10.8 $M=8$; 9.3 $M=16$; 8.2
M-ary QAM	$s_i(t)\sqrt{\frac{2}{T_s}}[A_i\cos\omega_c t + B_i\sin\omega_c t]$ $A_i, B_i = \pm a, \pm 3a, \ldots, \pm(\sqrt{M}-1)a$	2	$0.5\log_2 M$	$\frac{4}{\log_2 M}\left(1-\frac{1}{\sqrt{M}}\right)Q\left(\sqrt{\frac{2a^2}{N_0}}\right)$ $a^2 = \frac{3(\log_2 M)E_b}{2(M-1)}$	$M=2, 4$; 10.54 $M=16$; 14.3 $M=64$; 18.5 $M=256$; 24
Noncoherent M-ary FSK	$s_i(t) = \sqrt{\frac{2E_s}{T_s}}\cos\{[\omega_c + (i-1)\Delta\omega]t + \theta\}$ $\Delta\omega = 2\pi/T_s$ for orthogonal; $\theta =$ uniform in $[0, 2\pi]$	M	$\frac{2\log_2 M}{M+3}$	$\sum_{k=1}^{M-1}\binom{M-1}{k}\frac{M(-1)^{k+1}}{2(M-1)(k+1)}$ $\times \exp\left(-\frac{k}{k+1}\,\frac{E_s}{N_0}\right)E_s = (\log_2 M)E_b$	$M=2$, 14.2 $M=4$; 11.3 $M=8$; 9.7 $M=16$; 8.7
M-ary DPSK[a]	$s_i(t) = \sqrt{\frac{2E_s}{T_s}}\cos\left[\omega_c t + \frac{2\pi(i-1)}{M} + \gamma\right]$ $i = 1, 2, \ldots, M$; $\gamma =$ random phase angle	2	$0.5\log_2 M$	$\approx \frac{2}{\log_2 M}\sqrt{\frac{1+\cos\left(\frac{\pi}{M}\right)}{2\cos\left(\frac{\pi}{M}\right)}}$ $\times\, Q\left(\sqrt{\left[1-\cos\left(\frac{\pi}{M}\right)\right]\frac{2E_s}{N_0}}\right)$	$M=2, 4$; 12.7 $M=8$; 17 $M=16$; 21.5 $M=32$; 25.5
$\pi/4$-DQPSK	See text	2	1	$\approx Q(\sqrt{1.1716E_b/N_0})$	12.81

$Q(x) = \int_x^\infty \frac{\exp(-u^2/2)}{\sqrt{2\pi}}\,du \approx \frac{\exp(-x^2/2)}{\sqrt{2\pi}x}$, $x \ge 3$ (less than 10% error);

[a] The exact bit error-probability for binary DPSK is $P_b = \frac{1}{2}\exp(-E_b/N_0)$.

scheme; T_b is the symbol duration for a binary scheme (bit period); T_s is the symbol duration for an M-ary scheme; P_b denotes bit error probability; P_s denotes symbol error probability; $d_1(t)$ and $d_2(t)$ denote data sequences having values of ± 1 in T_s-second intervals; $\omega_c(f_c)$ denotes carrier frequency or nominal center frequency in rad/s (Hz); $\Delta\omega$ denotes frequency shift in rad/s; and M denotes the number of possible transmitted symbols in T_x seconds. In Table II, the signal set in each case is defined for the interval $0 \le t \le T_s$; it is understood that the symbols follow in sequence so that the ith signaling interval is $(i-1)T_s \le t \le iT_s$. To convert between symbol and bit energy, note that

$$E_s = (\log_2 M)E_b \tag{28}$$

The bit error probability can be related to the symbol error probability in one of two ways. If the procedure known as Gray encoding is used to ensure that an error in mistaking a symbol for its nearest neighbor results in only one bit error, then

$$P_b \approx P_s/\log_2 M \tag{29}$$

If orthogonal signal sets with equidistant signal points are used (e.g., M-ary FSK), then

$$P_b \approx \frac{M}{2(M-1)}P_s \tag{30}$$

Some observations about the results in Table II are as follows. Quadrature phase-shift keying (QPSK) and offset quadrature phase-shift keying (OQPSK) offer a good compromise between bandwidth efficiency and error-rate performance, with MSK following close behind. M-ary PSK allows an increase in bandwidth efficiency with M at the expense of error-rate performance, and this increase in bandwidth efficiency goes only as the logarithm of M. M-ary FSK allows better error-rate performance with increasing M at the expense of a decrease in bandwidth efficiency. Another modulation scheme not shown, known as vertices-of-a-hypercube, allows essentially an even trade-off between these two performance measures. Not addressed in Table II are other considerations, such as complexity of hardware and constant versus nonconstant envelope, to name only two such considerations.

D. Pulse Shaping

In the digital signaling schemes discussed so far, it has tacitly been assumed that the pulse shapes are rectangular, which implies an infinite bandwidth. It is possible to use pulse shapes with limited bandwidth and not cause intersymbol interference (ISI) due to the bandlimiting. Assume an antipodal binary signaling scheme wherein the bits are represented by signed impulses. The communication system consisting of a cascade of transmitter, channel, and receiver filters with a sampler-thresholder operation following the receiver filter for decision making. If overall pulse response spectrum of the system, $P_r(f)$, satisfies

$$\sum_{k=-\infty}^{\infty} P_r\left(f + \frac{k}{T_b}\right) = \text{constant}, \quad |f| \le \frac{1}{2T_b} \tag{31}$$

it can be proved that the samples at the receiver filter output will exhibit the zero-ISI property

$$p_r(nT_b) = \begin{cases} 1, n = 0 \\ 0, n \ne 0 \end{cases} \tag{32}$$

This is known as Nyquist's pulse-shaping criterion. A pulse spectrum whose inverse Fourier transform has the property (32) is the raised cosine spectrum given by

$$P_{RC}(f) = \begin{cases} T_b, & 0 \le |f| \le \dfrac{1-\beta}{2T_b} \\ \dfrac{T_b}{2}\left\{1 + \cos\left[\dfrac{\pi T_b}{\beta}\left(|f| - \dfrac{1-\beta}{2T_b}\right)\right]\right\}, & \dfrac{1-\beta}{2T_b} \le |f| \le \dfrac{1+\beta}{2T_b} \\ 0, & |f| > \dfrac{1+\beta}{2T_b} \end{cases} \tag{33}$$

with corresponding time-domain pulse shape family

$$p_{RC}(t) = \frac{\cos(\pi\beta t/T_b)}{1-(2\beta t/T_b)^2}\operatorname{sinc}(t/T_b) \tag{34}$$

Both are shown in Fig. 5.

It can be shown that the optimum transmitter and receiver filter frequency responses, $H_R(f)$ and $H_T(f)$, respectively, required to maintain the zero ISI property and to minimize the probability of detection error are given by

$$|H_R(f)| = \frac{\alpha|P_r(f)|^{1/2}}{G_n^{1/4}(f)|H_C(f)|^{1/2}} \tag{35}$$

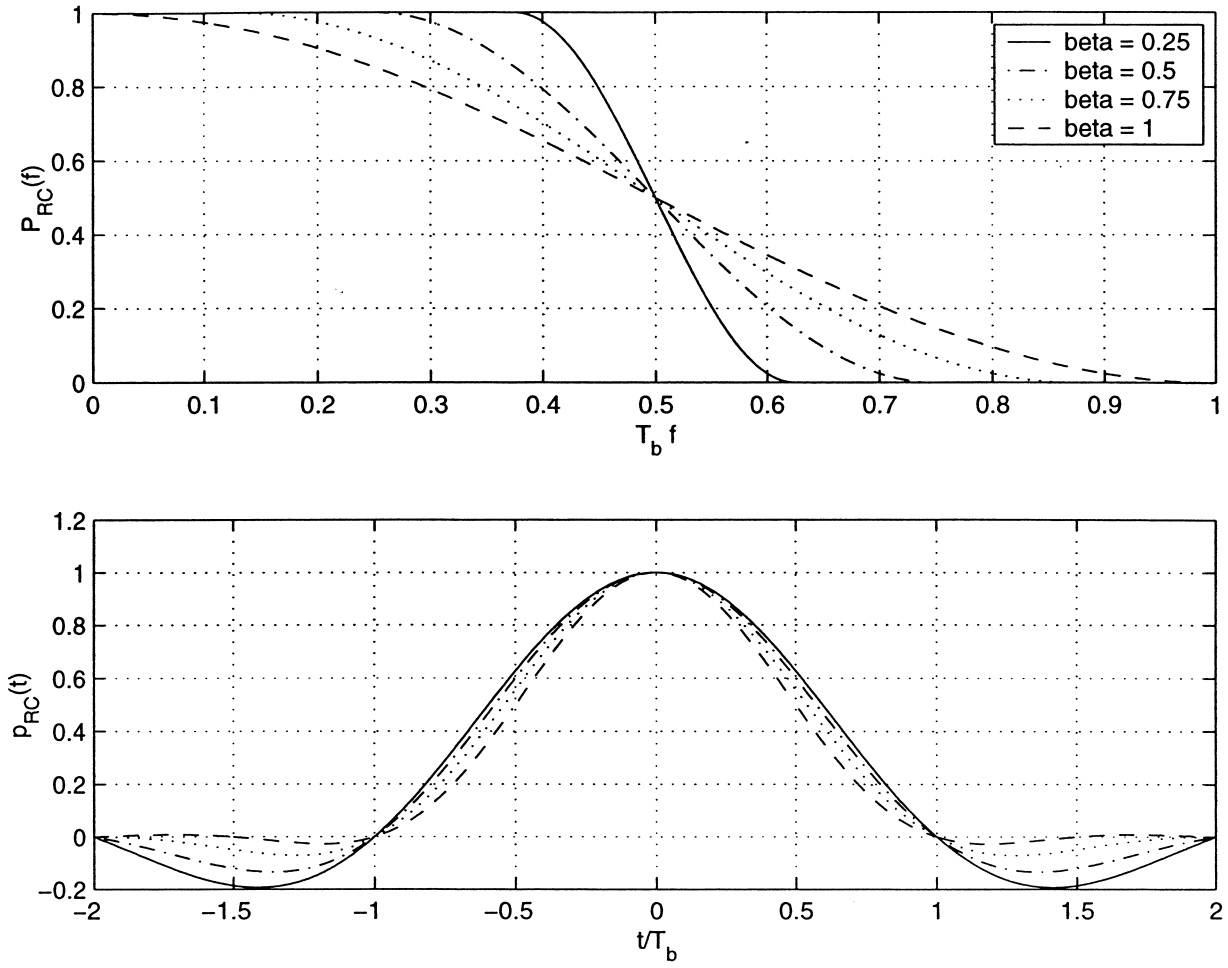


FIGURE 5 Raised-cosine spectra and pulses.

and

$$|H_{\mathrm{T}}(f)| = \frac{\beta |P_{\mathrm{r}}(f)|^{1/2} G_{\mathrm{n}}^{1/4}(f)}{|H_{\mathrm{C}}(f)|^{1/2}} \tag{36}$$

where α and β are arbitrary constants, $P_{\mathrm{r}}(f)$ is the receive pulse spectrum (to have the zero ISI property), $G_{\mathrm{n}}(f)$ is the additive noise spectrum, and $H_{\mathrm{C}}(f)$ is the channel frequency response. If $G_{\mathrm{n}}(f) = N_0/2$ (white noise), $H_{\mathrm{R}}(f)$ and $H_{\mathrm{T}}(f)$ are proportional to $|P_{\mathrm{r}}(f)|^{1/2}$, which gives rise to the often-heard term "square-root-raised cosine" pulse shaping.

E. Noncoherent Digital Modulation Schemes

A noncoherent digital modulation scheme is one not requiring a reference signal at the de-modulator that is in phase coherence with the received carrier. Two examples of such schemes are differential phase-shift keying (DPSK) and noncoherent frequency-shift keying (NFSK). These are discussed briefly in the following.

1. DPSK

In DPSK, the previous symbol period is used to provide a phase reference for the current symbol period. This procedure presupposes that phase perturbations introduced in the channel vary slowly compared with a symbol period and that there is a known phase relationship between adjacent bits. The latter is provided by using a procedure known as differential encoding on the data before PSK modulation of the carrier. Although differential encoding can be carried out for M-ary alphabets, it is easiest toexplain it in terms of binary data. Let the data sequence from a binary source be denoted as $\{D_n\}$ and the differentially encoded data sequence be denoted as $\{C_n\}$. The operation used to produce the nth bit of the differentially encoded sequence is

$$C_n = D_n \oplus C_{n-1} \tag{37}$$

where $\oplus$ denotes a binary add without carry (also called an exclusive-OR). For example, the sequence 10111001010

is differentially encoded as 1̲00101110011, where the first bit is an arbitrarily selected reference bit. The differentially encoded sequence is then used to PSK modulate a carrier, which results in the DPSK modulated signal.

To generalize the concept of binary DPSK to the case of M-ary DPSK, let the transmitted carrier phase angle at symbol time $n-1$ be α_{n-1} and suppose that it is desired to transmit symbol $\beta_n = \Phi$ at time n, where β_n takes on any of the values given by

$$\beta_n = \frac{(2l-1)\pi}{M}, \quad l = 1, 2, \ldots, M \tag{38}$$

The *transmitted phase* at time n is then

$$\alpha_n = \alpha_{n-1} + \phi \tag{39}$$

Suppose that the phases received corresponding to α_{n-1} and α_n are $\theta_{n-1} = \alpha_{n-1} + \gamma$ and $\theta_n = \alpha_n + \gamma$, respectively, where γ is the unknown phase shift introduced by the channel. A correct decision is made at the receiver whenthe *received phase difference* $\theta = \theta_n - \theta_{n-1}$ is such that

$$\phi - \frac{\pi}{M} < \theta \leq \phi + \frac{\pi}{M} \tag{40}$$

The decision rule (40) is identical to that used for coherent M-ary PSK, except now the *phase differences* between successive symbols are used in the decision procedure; i.e., into which $2\pi/M$ wedge does the received phase difference fall?

2. NFSK

Noncoherent FSK uses frequency shifts, as for coherent FSK, to represent the M possible signals selected from each signaling interval, but phase coherent references are not used for detection at the receiver. Consequently, the minimum frequency separation between signals must be somewhat greater than for coherent FSK, typically $2/T_s$. The optimum detector now involves correlation with a sine and cosine reference for each possible transmitted signal (accurate frequency references assumed) because of the unknown received phase.

F. Special Binary Modulation Schemes

Two modulation schemes of importance in cellular radio systems are described next. These are Gaussian minimum-shift keying (GMSK) and $\pi/4$-differential phase-shift keying ($\pi/4$-DQPSK).

1. Gaussian MSK

This scheme is implemented by passing the bipolar bit stream through a low-pass filter with Gaussian impulse response (and therefore Gaussian transfer function) and passing this filtered bit stream into a voltage controlled oscillator (VCO). The output of the VCO is then the GMSK-modulated signal. An advantage of this modulation scheme is that it has much lower out-of-band power than ordinary MSK. Furthermore, it can be coherently detected, differentially detected, or discriminator detected.

In Table II, $\text{erf}(u) = 1 - \text{erfc}(u) = 1 - 2Q(\sqrt{2}u)$. The other parameters are defined as follows: P_c is the total power in the modulated signal; f_c is the carrier frequency; f_d is the frequency deviation; B is the Gaussian filter bandwidth; T_b is the bit period; $g(t)$ is the phase pulse shape function. Typical waveforms for GMSK are shown in Fig. 6. The bit error probability is about 3 dB worse than BPSK for $BT_b = 0.5$ and about 4 dB worse for $BT_b = 0.2$. Its bandwidth occupancy has been characterized through computer simulation (Morota and Haride, 1981) and is given in Table III.

2. $\pi/4$-Differential QPSK ($\pi/4$-DQPSK)

$\pi/4$-DQPSK is an eight-phase modulation scheme, but since successive symbols cannot use adjacent phases, it can be thought of as two DQPSK modulation schemes, where a given symbol modulates one set of quadrature carriers, and the next symbol modulates another set of quadrature carriers phase shifted by $\pi/4$ radians from the first set. The quadrature channel symbols, I_k and Q_k, in a given time slot are related to those in the immediately preceding time slot by the equations

$$\begin{aligned} I_k &= I_{k-1}\cos\Delta\phi_k - Q_{k-1}\sin\Delta_k; \\ Q_k &= I_{k-1}\sin\Delta\phi_k + Q_{k-1}\cos\Delta_k \end{aligned} \tag{41}$$

The phase difference from time slot $k-1$ to k is determined by the quadrature-channel symbol streams, d_{2k-1} and d_{2k}, which are derived as the even- and odd-indexed bits of the input data to the modulator. The mapping is as given in Table IV.

The advantages of this modulation scheme are that the symbol transitions never go through the origin, which gives lower envelop deviation when the modulated signal is filtered, and there is always a transition between symbols, allowing easier synchronization the demodulator.

G. Other Considerations

Much is left unsaid in this synopsis of modulation theory. An important untouched topic is combined modulation and coding. Previously, coding was treated as an add-on to modulation for providing a trade between bandwidth and power efficiency. In the late 1970s it was confirmed that the use of combined coding and modulation gives simultaneous inproved power and bandwidth efficiency. One such scheme for doing this trellis-coded modulation (Ungerboeck, 1987).

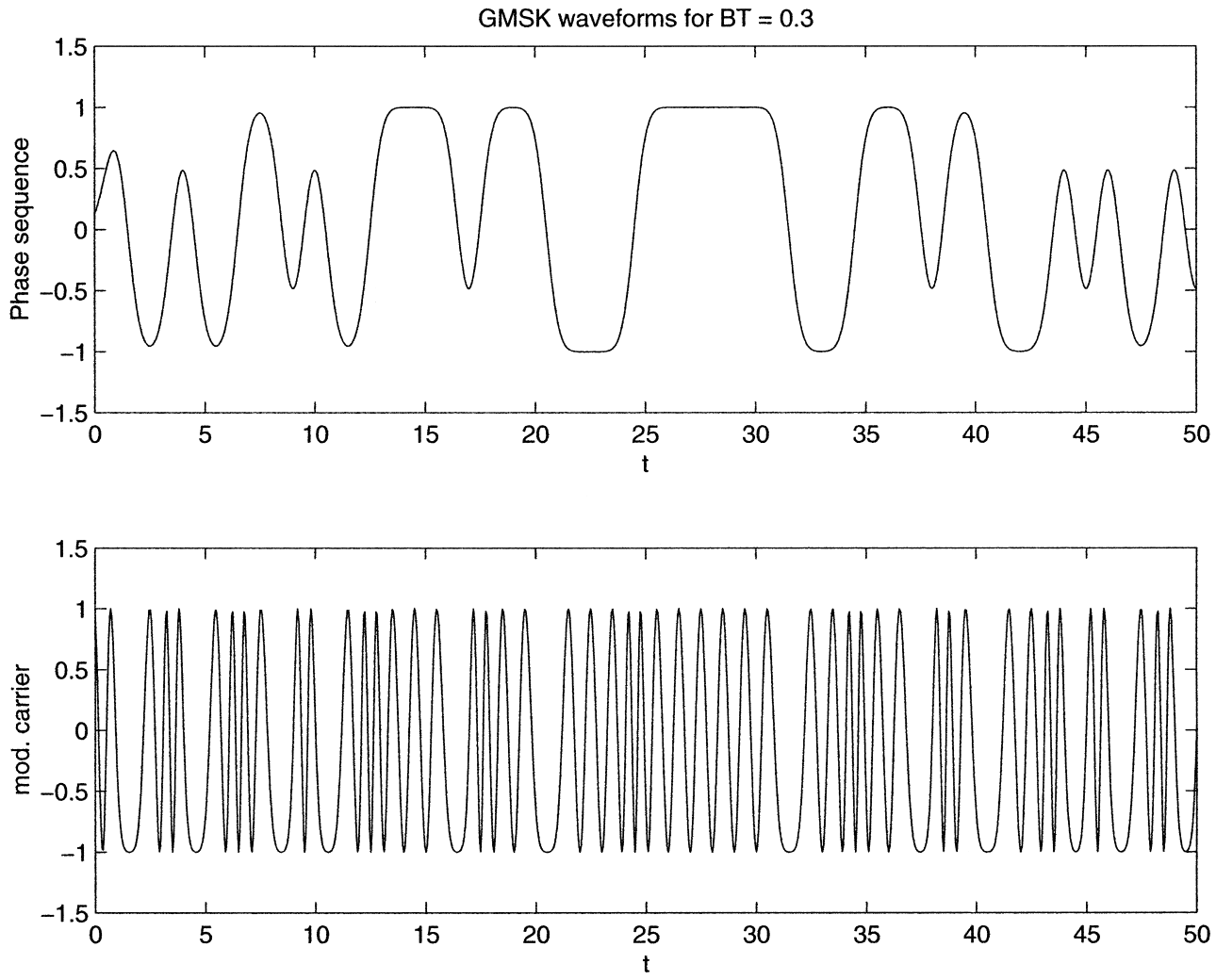


FIGURE 6 GMSK waveforms; $BT_b = 0.3$.

Spread spectrum modulation can provide performance improvement in jamming and multipath environments. Such schemes include any modulation scheme in which the bandwidth of the modulated signal is spread well beyond that of the modulating signal itself, independently of the modulating signal bandwidth. Two Primary means of achieving spread spectrum are direct sequence spreading and frequency hop spreading. Spread spectrum also provides a means for multiple access to a common communication resource, such as a satellite repeater or a cellular radio base station (Peterson, Ziemer, and Borth, 1995).

Finally, we have given communication system performance only in the most benign environment of AWGN. Performance in fading/multipath channels and means for combating their detrimental effects is also of importance, particularly for cellular radio (Proakis, 1995).

TABLE III Occupied Bandwidth × T_b Containing a Given Percent Power for GMSK

% Power → BT_b↓	90	99	99.9	99.99
0.2	0.52	0.79	0.99	1.22
0.25	0.57	0.86	1.09	1.37
0.5	0.69	1.04	1.33	2.08
MSK	0.78	1.20	2.76	6.00

TABLE IV Mapping from Bits to Phase for $\pi/4$-DQPSK

Odd bit, d_{2k-1}	Even bit, d_{2k}	$\Delta\phi_k$
0	0	$\pi/4$
0	1	$-\pi/4$
1	0	$3\pi/4$
1	1	$-3\pi/4$

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SYSTEMS, CIVILIAN • INFORMATION THEORY • MULTIPLEXING • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL • SPREAD-SPECTRUM SYSTEMS

BIBLIOGRAPHY

Berrou, C., Glavieux, A., and Thitimajshima, P. (1993). Near Shannon limit error-correcting coding and decoding: turbo codes, *IEEE Inter. Commun. Conf. Record*, pp. 1064–1070, June 1993.

Morota, K., and Haride, K. (1981). GMSK modulation for digital mobile radio telephony, *IEEE Trans. Commun*. **COM-29,** 1044–1050.

Pawula, R. (1984). Asympotics and error rate bounds for M-ary DPSK, *IEEE Trans. Commun*. **COM-32,** 93–94.

Peterson, R. L., Ziemer, R. E., and Borth, D. A. (1995). "Introduction to Spread Spectrum Communications," Prentice Hall, Upper Saddle River, NJ.

Proakis, J. G. (1995). "Digital Communications," McGraw Hall, New York.

Ungerboeck, G. (1982). Channel coding with multilevel/phase signals, *IEEE Trans. Inform. Theory* **IT-28,** 55–66.

Ungerboek, G. (1987). Trellis-coded modulation with redundant signal sets, Part I: Introduction, *IEEE Commun. Mag*. **25,** 5–11; Part II: State of the art, *IEEE Commun. Mag.* **25,** 12–21.

Ziemer, R. E., and Tranter, W. H. (1995). "Principles of Communications: Systems, Modulation, and Noise," 4th ed. Willey, New York.



Multiplexing

Rodger Ziemer
University of Colorado at Colorado Springs

I. Multiplexing of Analog Sources
II. Time-Division Multiplexing of Digital Data
III. Multiple Access Techniques

GLOSSARY

Aloha A random access scheme wherein a user desiring to use the communication medium simply sends its packet and waits for an acknowledgment. If none is received after a time-out period, the user attempts the transmission again.

Asynchronous transfer mode (ATM) A packetized transmission technique that makes use of 53-byte packets.

Asynchronous multiplexing Sharing of a common channel resource by several message sources using time division multiplexing whereby the use of a common clock for all sources is not required.

Binary digit (bit) Symbol that conveys one of two possible messages through a communications system.

Bit (symbol) stuffing A technique for bringing nearly synchronous bit (symbol) streams into synchronism by inserting bits and removing them at the information sink.

Byte An eight-bit word.

Character (word) Sequence of digital symbols (bits) that conveys a particular piece of information (e.g., letter of the English alphabet) in addition to including certain overhead information such as start bits, stop bits, and parity check bits.

Carrier-sense multiple access (CSMA) A random access scheme wherein a user first listens for activity on the transmission medium and sends a packet only when it hears none.

Circuit switching Method of information transmission that sets up a circuit before the call and holds the circuit for the entire call.

Code-division multiple access (CDMA) Sharing of remote channel resources through the assignment of a unique code to each user, which is used to encode the information transmitted by that user. This encoding is usually done through a phase modulation process of each bit using the code or through using the code to frequency hop the carrier in a unique fashion for each user.

Demand assigned multiple access (DAMA) Multiple access system whereby communication resources are assigned as demanded by the users.

Fixed assigned multiple access (FAMA) Multiple access system whereby communication resources are assigned on a fixed basis to the users.

Frame Sequence of data from several different sources. The information can be interlaced on a bit-by-bit basis or on a character-by-character basis and often includes overhead information for synchronization purposes, error correction purposes, etc.

Frequency-division multiple access (FDMA) Remote sharing of communication channel resources through allocation in frequency in a similar fashion to FDM. The assignment of these resources may be fixed or varied with demand.

Frequency-division multiplexing (FDM) Scheme that uses common communication channel resources by fixed allocation in frequency.

Internet A scaleable network of networks that has decentralized control and uses store-and-forward means for sending messages from one user to another.

Internet protocol (IP) The protocol used for sending messages from one user to another on the Internet.

Multiple access Method that allows the use of a remote communication resource such as a relay satellite, by several different users.

Packet switching Method of information transmission wherein the information is sent in packets with a payload and address portion. The packet is routed through the network by utilizing the address information.

Polling Multiple access scheme where a central controller polls each user in turn to determine if it has information to send.

Quadrature multiplexing A multiplexing scheme that carries two message signals in the same bandwidth by modulating one on a carrier and the second on a carrier at the same frequency in phase quadrature with the first.

Slot Portion of the total time in a frame assigned to a given user in a time division multiple access system.

Store-and-forward The procedure of relaying information through a network by temporarily storing packets in buffered switching nodes and relaying them using the packet address information once a link on the way to the destination becomes available.

Synchronous In reference to a time division multiplexed system, this term implies that all sources are kept in step from the same clock.

Time-division multiple access (TDMA) Sharing of a remote communication resource, such as a relay satellite, on a time division basis.

Time-division multiplexing (TDM) Scheme that uses common communication channel resources by allocation on a time interval basis.

Token passing Multiple access scheme where a given user gains access to the communications medium by being passed a token. The token is passed from user to user in round-robin fashion.

MULTIPLEXING is the use of some means of interleaving narrow-band or slow-speed data from multiple sources to make use of a wide-band or high-speed transmission resource (channel). Depending on whether a master system clock is used or not, multiplexing is either synchronous or asynchronous. Closely related to multiplexing is the idea of multiple access. Multiplexing presupposes fixed assignment of channel resources at a local level to the separate users, whereas multiple access schemes make use of a remote channel resource, such as a communications satellite, with a possibly dynamic resource assignment. The two most common means of multiplexing are time division multiplexing (TDM) and frequency division multiplexing (FDM). The former interleaves data from separate sources in time, whereas the latter does so in frequency. When applied to multiple access, these respective methods are referred to as time division multiple access (TDMA) and frequency division multiple access (FDMA), respectively. There is a third method for multiple access, called code division multiple access (CDMA), that makes use of a unique code assigned to each user in order to separate each user's transmission from the others. Other access schemes described briefly in the sequel are Aloha, carrier-sense multiple access, polling, and token passing. Also discussed and compared are the concepts of circuit switching and store-and-forward transmission.

I. MULTIPLEXING OF ANALOG SOURCES

A. Frequency Division Multiplexing

In the article entitled "Modulation," it was shown that multiplication of a low-pass signal by a sinusoid of f_0 Hz translates the spectrum of the signal to f_0 Hz. This principle can be used to advantage when it is desired to transmit several low-pass message signals through the same channel. A frequency division multiplexed transmitter for n messages, $m_1(t), m_2(t), \ldots, m_n(t)$, with bandwidths $W_1, W_2, \ldots, W_n$, respectively, is shown in block diagram form in Fig. 1a. Each message is used to modulate a subcarrier, assumed in this case to be single-sideband. The outputs of the subcarrier modulators are summed and possibly used to modulate a carrier in order to move the entire spectrum to some preassigned frequency location. This step may not be required in some applications.

It is assumed that the subcarrier frequencies are separated such that $f_2 - f_2 \geq W_1$, $f_3 - f_2 \geq W_2, \ldots, f_n - f_{n-1} \geq W_{n-1}$ so that the separate spread-spectrum modulated message signal spectra are not overlapping. It is clear that if single-sideband subcarrier modulation with the minimum separation between subcarriers is used, the minimum transmission bandwidth required is given

$$B_T = \sum_{i=1}^{n} W_i \tag{1}$$

The demodulation of FDM at the receiving end, as shown in Fig. 1b, is accomplished by first going through

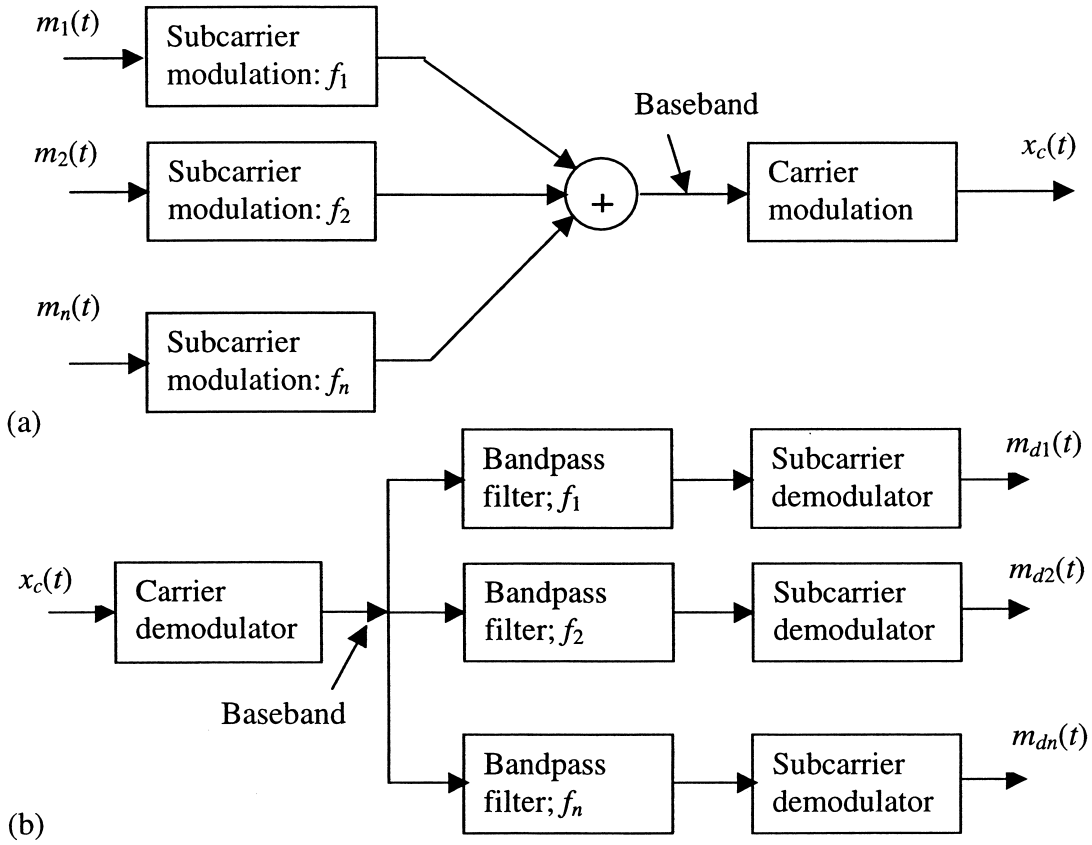


FIGURE 1 (a) An FDM modulator and (b) a demodulator.

a carrier demodulator, if carrier modulation was used at the transmitting end, and subsequently through a bank of bandpass filters centered on the appropriate subcarrier-modulated signals in order to separate them from each other. The final step is to pass each subcarrier-modulated signal through a subcarrier demodulator. In this case, coherent demodulation is used because of the single-sideband modulation used at the transmitting end. A coherent carrier reference is required, and a convenient way to achieve this is by sending it through one of the subchannels, usually the highest frequency one, and then frequency dividing it down to the required subcarrier frequencies.

An example of a current FDM application is wideband data delivery to the home or small business. Many approaches to this so-called last-mile challenge are being pursued. One of these is fuller utilization of the cable television plant, wherein upstream digital data channels are added to the normal analog cable TV spectrum (50–450 MHz) below 50 MHz and downstream data channels are added above 450 MHz (450–750 MHz). Within each of these FDM frequency channels, actual delivery of the data is made in a time-multiplexed fashion (see the discussion on TDM, below) as determined by a bandwidth manager.

Frequency division multiplexing is a simple scheme to implement. Its major disadvantage is that any channel nonlinearities will generate harmonic (integer multiples of the subcarrier frequencies) and intermodulation (sums and differences of harmonics of the subcarrier frequencies) distortion that may interfere with the desired messages.

B. Time-Division Multiplexing

Time-division multiplexing can be applied in instances where the messages are represented in pulse modulation format. Figure 2 illustrates such a scheme, where pulse amplitude modulation is used to represent the messages $m_1(t), m_2(t), \ldots, m_n(t)$, which are now assumed to all be of bandwidth W Hz. A commutator samples each message in turn, which according the to sampling theorem must be at a minimum rate of $2W$ samples per second per message. At the receiving end, a second commutator, which is synchronized with the one at the transmitting end, deinterleaves the samples corresponding to the respective messages. A low-pass filter then allows the recovery of each message from its sample values. Using the sampling theorem in reverse, it can be shown that the theoretical

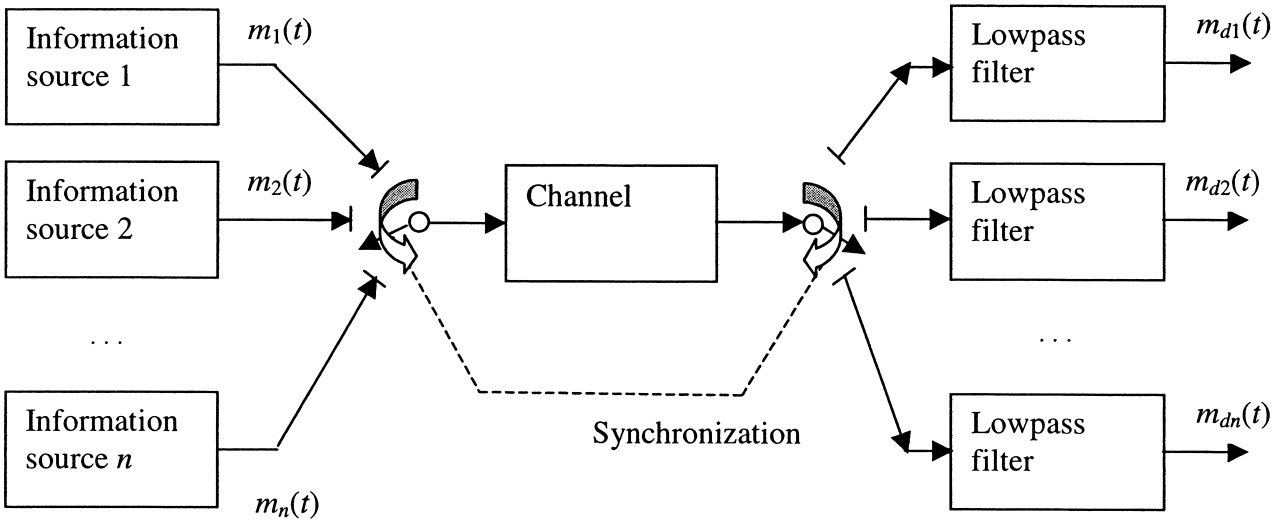


FIGURE 2 A time-division multiplexed system.

minimum transmission bandwidth requirement for TDM is exactly the same as that for FDM.

Any analog pulse modulation format can be used in TDM, and in fact, there is no restriction to the pulses representing analog messages. However, additional considerations, such as framing, are necessary when the messages are digital (to be discussed later). Time-division multiplexing avoids the crosstalk problems of FDM. However, it is necessary to maintain synchronization of all message samples and to synchronize the decommutation at the receiving end with the commutation at the transmitting end. Time-division multiplexing can also accommodate messages of unequal bandwidths through the use of creative sampling. For example, for three messages m_1, m_2, and m_3 of bandwidths W, W, and $2W$, the commutator would consist of one contact for m_1, one for m_2, and two for m_3 in the sequence $m_1 m_3 m_2 m_3 m_1 m_3 m_2 m_3 \ldots$.

C. Quadrature Multiplexing

A technique for conveying two different message signals at the same carrier frequency is to modulate one onto a cosine carrier, say of frequency f_c, and the other onto a sine carrier of the same frequency. This technique is referred to as quadrature multiplexing. If the two message signals are denoted as $m_1(t)$ and $m_2(t)$, the quadrature-multiplexed signal is given by

$$x_c(t) = A_c[m_1(t)\cos 2\pi f_c t + m_2(t)\sin 2\pi f_c t] \qquad (2)$$

The individual message signals, $m_1(t)$ and $m_2(t)$, are recovered by multiplying the received signal by $2\cos 2\pi f_c t$ and $2\sin 2\pi f_c t$, and passing these products through low-pass filters of bandwidth at least as large as the bandwidths of $m_1(t)$ and $m_2(t)$, with the resultant outputs of the low-pass filters being $m_1(t)$ and $m_2(t)$, respectively. An example of a system employing quadrature multiplexing is the Global Positioning Satellite system, where the clear/acquisition (C/A) code is transmitted on one quadrature carrier and the precise (P) code is transmitted on the other.

II. TIME-DIVISION MULTIPLEXING OF DIGITAL DATA

A. Introduction

Digital data can originate in two ways: (1) it may come from a source that is inherently digital, such as a computer or messages consisting of written text, or (2) it can result from sampling an analog message, such as voice. In the latter case, an intermediate step known as analog-to-digital (A/D) conversion is necessary, wherein the analog message is sampled in accordance with the sampling theorem and the samples converted or encoded into digital words. Encoding is usually binary and can be accomplished with any of several different types of encoders (called vocoders). It is not the purpose of this discussion to enlarge upon vocoder types. When analog signals are sampled and encoded into digital format using an A/D converter, the process is referred to as pulse code modulation (PCM). It is often necessary to time division multiplex PCM signals to fully utilize the channel. Such time division multiplexing of PCM data can be accomplished in various ways as discussed next.

B. Synchronous Time-Division Multiplexing of PCM data

In synchronous multiplexing of PCM data, it is assumed that the rates of all sources are derived from a common

clock. This guarantees high efficiency, but at the expense of greater complexity.

An advantage of a synchronous TDM PCM system is that it is completely feasible to intermix analog and digital sources. For example, consider the time-division multiplexing of two 2-kHz analog sources, a 4-kHz analog source, and eight 8-kb/s digital sources, with 4-bit representation used for each analog sample. By virtue of the sampling theorem, each analog source must be sampled at a rate at least twice its bandwidth. If the analog sources are sampled at their minimum rates of 4, 4, and 8 kHz, respectively, the samples combined, and the combined samples converted to 8-bit digital format, then the rate of samples taken from the analog sources is 128 kb/s. These bits, representing the analog sources, can then be interlaced with the bits from the digital sources to produce a combined bit stream of 192 kb/s. Letting a_j represent the jth bit from the combined outputs of the analog sources and d_{ij} represent the jth bit from the ith digital source, the bit stream would be $a_1 d_{11} a_2 d_{21} a_3 d_{31} a_4 d_{41} a_5 d_{51} a_6 d_{61} a_7 d_{71} a_8 d_{81} a_9 d_{12} \ldots$. This example illustrates bit-by-bit TDM. If the various bit streams are not synchronous, but nearly so, a technique known as pulse stuffing can be used to make them synchronous. This will be discussed later.

An alternative technique to this example is to interlace words or characters from different sources. A complete sequence of characters along with overhead bits for purposes such as synchronization and error correction constitute a frame of data. For example, in the T1 carrier system for TDM telephone transmission in the United States, each so-called DS1 line consists of 24 channels, each channel consisting of 8-bit words representing samples taken at 8-kHz rates [one sample per 125 microseconds (μs)] from each telephone conversation. Thus, in each frame there are 192 bits representing telephone conversations plus 1 bit added for framing. If b_j represents the jth bit in the ith channel and f represents the framing bit, the frame sequence is $f b_{11} b_{12} b_{13} b_{14} b_{15} b_{16} b_{17} b_{18} b_{21} b_{22} b_{23} b_{24} b_{25} b_{26} b_{27} b_{28} \ldots b_{24,8}$. Each bit is (125 μs)/(193 bits/frame) = 0.6477 μs in duration and is represented in bipolar format (A volts for half the bit duration, and $-A$ volts for the other half) so that no DC component exists on the line. The framing bit for the even-numbered frames follows the sequence 001110, and for the odd-numbered frames, it follows the sequence 101010. This is so that the signaling information, which consists of the eighth bit of each of the 24 channels on every sixth frame, can be identified. The data rate corresponding to a 0.6477-μs bit time is 1.544 Mb/s, requiring a minimum channel bandwidth of 1.544 MHz. When compared with the bandwidth requirements of a 24-channel FDM system, it is seen that TDM requires much more—roughly a factor of 11 more. Yet, the flexibility offered by TDM systems and their resistance to nonlinear distortion makes them preferable to FDM systems. With the advent of fiber-optic transmission media, TDM is an even more attractive medium for voice and data transmission.

Moving up the hierarchy, four DS1 lines are combined to form a DS2 stream at 6.312 Mb/s, seven DS2 streams are combined to form a DS3 stream at a total rate of 44.74 Mb/s, and six DS3 streams are combined to form a DS4 stream at 274 Mb/s.

Although telephone systems extensively make use of TDM, they are not slaved to a master clock for a number of reasons, one of which is that temperature variations of cables mean that the number of symbols in a cable vary over time as result of propagation velocity changes. Clocks could be exactly synchronous at all sources, and yet symbol streams may be out of synchronism because of cable temperature variations.

A critical device in making multiplexing work in a practical environment where symbols stream rates vary slightly from the desired average rate is an *elastic store*. This can be visualized as a circular buffer whose input (or write) clock is controlled by the incoming data clock and whose output (or read) clock is controlled by the outgoing data clock. There is one for each multiplexed symbol stream and one for each demultiplexed symbol stream. Ideally, the read and write clocks are the same on average, and symbols do not build up or become depleted in the buffer.

One method to reclock symbols in nearly synchronous symbol streams is by *symbol* or *bit stuffing*. This consists of adding symbols to slower symbol streams at predetermined places in order to bring all incoming streams to a multiplexer up to the same average rate. At the demultiplexer, the stuffed symbols are removed since their location is known. The example below, where $\times$ denotes data and $\bigcirc$ denotes stuff symbol, shows how this works:

$\frac{4}{5}$ rate stream: $\times\times\times\times\bigcirc\times\times\times\times\bigcirc\times\times\times\times\bigcirc\times\times\cdots$

$\frac{3}{4}$ rate stream: $\times\times\times\bigcirc\times\times\times\bigcirc\times\times\times\bigcirc\times\times\times\bigcirc\times\cdots$

$\frac{2}{3}$ rate stream: $\times\times\bigcirc\times\times\bigcirc\times\times\bigcirc\times\times\bigcirc\times\times\bigcirc\times\times\cdots$

Bit stuffing has another use, which is to ensure that a frame synchronization sequence is not accidentally duplicated by a like pattern of data bits. In an n-bit frame synchronization sequence, if $n-1$ data bits match, the nth bit is stuffed to ensure that the synchronization sequence is not duplicated.

C. Asynchronous Time Division Multiplexing oF PCM Data

For asynchronous multiplexing, the information bits are sent in blocks, called characters or words. Each character begins with a start bit followed by several information bits and finally by one or two stop bits. The receiver clock starts with the arrival of the start bit. As a result, no synchronization with a master clock is necessary. Keyboard terminals are typical asynchronous sources and operate with a 7-bit ASCII code with each character having appended to it a start bit, a parity bit, and one stop bit for a total character length of 10 bits. In asynchronous multiplexing, the data from multiple sources is interleaved on a character-by-character basis instead of on a bit-by-bit basis. Because of the required start and stop bits, asynchronous multiplexing is less efficient than synchronous multiplexing, although distribution of a clock is required for the latter.

D. Circuit Switching versus Packet Switching

The telephone network carries mostly voice with some data, consisting of primarily of modem-linked computers and facsimile. Because telephone networks originated for the purpose of carrying voice traffic, constraints are placed on them in terms of bandwidth and maximum tolerable delay. As such, when a telephone conversation is set up, a fixed circuit is dedicated for the duration of that conversation. Thus, we say that the telephone network is *circuit switched*. In terms of TDM, for example, a portion of each frame is dedicated to a given user throughout the duration of a call whether conversation is actually present or not. On the other hand, if a computer is accessing data over a telephone-line hookup, this dedication of resources is not fully utilized at all times during the circuit hookup because of the bursty character of typical data transfers. For this reason, the Internet and other networks intended for providing data transfer are based on packet transfer. A packet has two parts—the *payload* (information content) and the *header* (for routing the packet). The header allows a form of data transfer known as *store-and-forward*, which is analogous to postal delivery of a letter.

Store-and-forward delivery has obvious advantages over the dedicated resource type of delivery typified by the telephone network. First, it is more economical, since the delivery of a packet is not required according to a specific schedule and therefore does not demand the dedicated resources required for a telephone call. Second, if part of the transmission resource is temporarily out of service, packets can be stored and delivered when service comes back. Disadvantages of packet-type delivery are potentially long delays in delivery of some packets and the necessity of buffers for temporary storage of packets while waiting for access to a given segment of the transmission medium. Also, if many users attempt to send data at the same time to the same destination, the buffers may become overloaded and some packets may be lost as a result. In contrast, data words will never be lost in a telephone network using dedicated resource transmission, such as TDM, because each occupies its particular time slot with no others contending for it.

A good example of a network for packet transfer is the Internet. Packet transfer on the Internet employs a two-part addressing hierarchy (called IP for *Internet Protocol*). The first part is the network number, which is globally unique, and the second part is the interface number, which is unique within a given network. Thus, each interface within the Internet is uniquely identified. Each Internet operator has a unique network number, and a given operator can then allocate globally unique IP addresses with that prefix to users on that network. Routers along a path from destination to user then relay packets by reading the destination address from an incoming packet's header and forwarding it with the aid of stored routing tables.

E. Asynchronous Transfer Mode

Although the Internet provides a good example of a widely used packetized transmission resource, it does have some potential flaws, among which are security and lack of ensured reliability. Another approach to countering the disadvantages of *Synchronous Transfer Mode* (STM) data transfer (e.g., the TDM resource allocation of telephone networks) is *Asynchronous Transfer Mode* (ATM). ATM makes use of a fixed-length packet consisting of a 48-byte (384 bits) payload and a 5-byte (40 bits) header. The 53-byte packet size is the result of a compromise in the ATM standardization process between the desire of U.S. telephone company representation to lower the overhead of the ATM packets and the desire of European/Japanese telephone companies to avoid the need for installing echo-canceling equipment. Voice transmission via ATM is accomplished by taking a block of digitized voice samples, adding a header, and handing the packet consisting to payload plus header to the network for transmission. The receiver implements the inverse process of that at the transmitter.

III. MULTIPLE ACCESS TECHNIQUES

Multiple access (MA), like multiplexing, involves sharing a communications resource between several users that broadcast their transmissions so that more than one other user may receive them. Examples of shared

communications resources include relay satellites, base stations in cellular radio, and local area networks. There are several techniques for utilizing a shared communications resource:

1. Frequency-division multiple access (FDMA)
2. Time-division multiple access (TDMA)
3. Code-division multiple access (CDMA)
4. Random access

The first two are analogous to FDM and TDM, respectively. CDMA utilizes a unique code assigned to each user. Random access takes various forms, two common ones being Aloha and carrier-sense multiple access (CSMA).

When using FDMA, TDMA, or CDMA, the communication resources may be utilized in a fixed assigned multiple access (FAMA) mode or in a demand-assigned multiple access (DAMA) mode. In the FAMA mode, the assignment format does not change even though the traffic

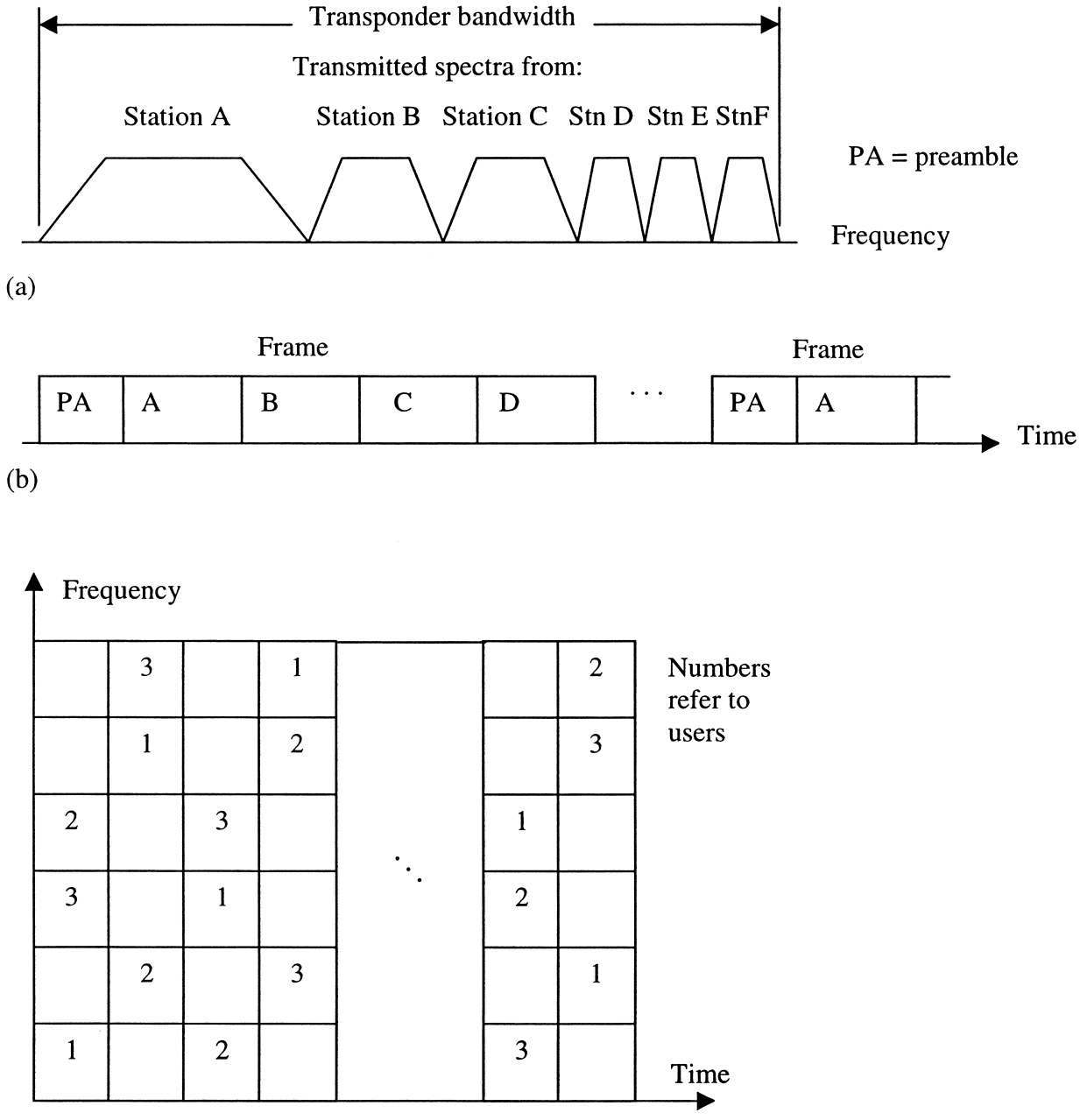


FIGURE 3 Illustrations of (a) FDMA, (b) TDMA, and (c) CDMA-FH systems. Numbers in (c) denote hopping sequences for channels 1, 2, and 3.

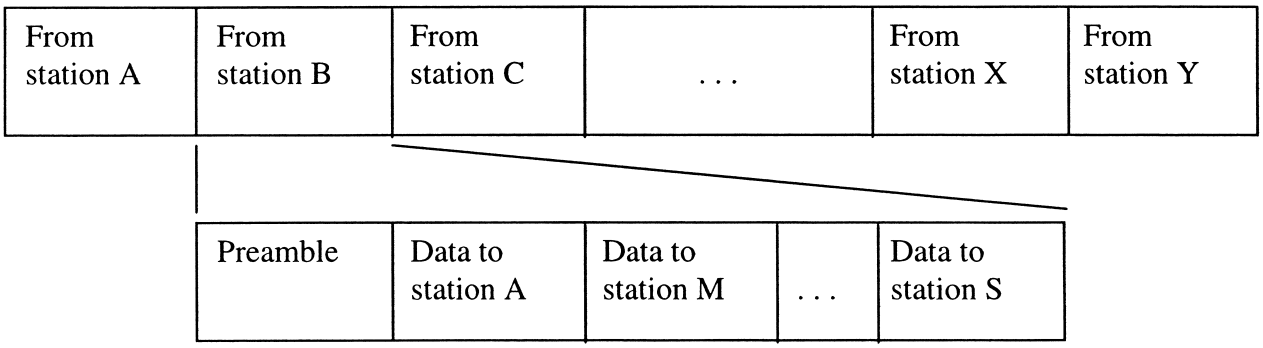


FIGURE 4 Detail of a TDMA frame format.

load of the various users may change. In the DAMA mode, the assignment format is adjusted, as needed depending on the traffic load of the various users.

Figure 3 illustrates the first three accessing schemes. In FDMA, signals from various users are stacked up in frequency, just as for FDM, as shown in Fig. 3a. Guard bands are maintained between adjacent signal spectra to minimize crosstalk between channels. If frequency slots are assigned permanently to the users, the system is referred to as FAMA. If some type of dynamic allocation scheme is used to assign frequency slots, it is referred to as a DAMA system.

In TDMA, the messages from various users are interlaced in time, just as for TDM, as shown in Fig. 3b. As illustrated in Fig. 4, the data from each user is conveyed in time intervals called slots. A number of slots make up a frame. Each slot is made up of a preamble plus information bits. The functions of the preamble are to provide identification and allow synchronization of the slot at the intended receiver. Guard times are utilized between each user's transmission to minimize crosstalk between channels. It is necessary to maintain overall network synchronization in TDMA.

In CDMA, each user is assigned a code that ideally does not correlate with the codes of other users. The codes of various users are then used to separate one user's transmissions from those of the other users. This can be accomplished in various fashions, one of which is to divide the time–frequency space occupied by all users into contiguous squares with one side representing a frequency increment and the other side representing a time increment, as shown in Fig. 3c. Each user then makes use of a unique sequence of squares in time and frequency to transmit its data as indicated by the patterns of integers shown in Fig. 3c. This technique is called frequency hop (FH) multiple access. Another technique is for each user to be assigned a unique spreading code and the ith user to send a 0 bit as $-c_i(t)$ and a 1 bit as $c_i(t)$, where $c_i(t)$ is its spreading code. This technique is known as direct-sequence spread-spectrum (DSSS) multiple access. Although CDMA schemes can be operated synchronously, they can be designed so as not to require synchronization. When operated without synchronization, one must account for multiple access noise, which is a manifestation of the partial correlation of a desired user's code with all other users' codes present on the system.

In the Aloha multiple access scheme, a user broadcasts an information packet and waits for an acknowledgment from the intended receiver. If none is received within a certain time interval, it is assumed that the packet was not received, perhaps because of collision with a packet sent from another user. In this event, the user retransmits after a random time interval. This process is continued until an acknowledgment is received, thus indicating a successful reception.

CSMA is similar to Aloha, except that a sending user first listens for activity on the communications medium (carrier sense). If it senses activity, it waits until activity is no longer sensed and then transmits. Of course, there may have been another user waiting also, in which case the two packets will collide. One way to avoid this almost sure collision situation is for both stations to back off a random time interval and then send. Another technique to avoid potential collisions is for a waiting user to flip a coin upon sensing no activity on the communications medium and send only if it comes up heads. Since each potential user performs the same procedure, there is a finite probability that only one station will send when the medium is sensed to be free.

There are variations of both Aloha and CSMA as described here. There are also other possible ways to utilize a communications resource, such as polling and token passing. As its name implies, polling sequentially polls all possible users to determine if they wish to send packets. An obvious disadvantage of such a procedure is that each user is polled whether it has something to send or not. The maximum delay for a given user to wait to send is the polling interval times the number of users. An obvious advantage is that each user gets its chance to use the medium. Token passing is similar, except that a token

is passed from user to user. When a given user has possession of the token, it has sole use of the communications medium.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SYSTEMS, CIVILIAN • INFORMATION THEORY • MODULATION

BIBLIOGRAPHY

Anderson, J. B. (1998). "Digital Transmission Engineering," Institute of Electrical and Electronics Engineers, Inc., New York.

Couch, L. W. (1996). "Digital and Analog Communications Systems," 4th ed., Wiley, New York.

Gagnaire, M. (1997). An overview of broad-band access technologies. *Proc. IEEE* **85,** 1958–1972.

Glover, I. A., and Grant, P. M. (1998). "Digital Communications," Prentice Hall, London.

IEEE Journal on Selected Areas in Communications (1988). Special Issue on Broadband Packet Communication, Vol. 6, No. 9.

IEEE Journal on Selected Areas in Communications (1991). Special Issue on Teletraffic Analysis of ATM Systems. Vol. 9, No. 3.

Keshav, S. (1997). "An Engineering Approach to Computer Networking," Addison-Wesley, Reading, MA.

Sklar, B. (1988). "Digital Communications: Fundamentals and Applications," Prentice-Hall, Englewood Cliffs, NJ.

Ziemer, R. E., and Peterson, R. L. (2000). "Introduction to Digital Communication," Prentice Hall, Upper Saddle River, NJ.

Ziemer, R. E., and Tranter, W. H. (1995). "Principles of Communications: Systems, Modulation, and Noise," 4th ed., Wiley, New York.



Networks for Data Communication

Paul S. Kreager
Washington State University

I. Network Basics
II. Network Intelligence
III. Network and Information Management
IV. Local-Area Networks
V. Wide-Area Networking

GLOSSARY

Asynchronous Method of synchronizing serial character-level data between sender and receiver, characterized by the addition of start and stop bits before and after the bitstring representing the character. The synchronizing bits are added before the sender transmits the information over the network, allowing the receiver to tell when the character begins and ends. Decoding of the character bits is done solely by the receiver, knowing the predetermined data transmission rate. Contrasts with synchronous data transmission.

Bandwidth Range of frequencies that a communication line can pass. In data communications, the term is also loosely used to denote the bit transmission rate (bits per second) that a line will pass.

Baseband Form of data communication where a communication line is turned over to the sole use of one subscriber and binary information is put directly on the line in its direct electrical current or voltage equivalent. Form of time-division multiplexing and generally in contrast to broadband communication.

Broadband Form of data communication where several subscribers share and can simultaneously use one common communication line. Each subscriber's data are modulated over a carrier frequency, i.e., information is frequency-division multiplexed. Generally, in contrast to baseband communication.

Carrier Company in the business of carrying data over its transmission-line facilities, e.g., telephone companies. Also, an electrical signal used to carry information by modulation of the signal. Modulating a fixed-frequency carrier amplitude with voice information in AM radio is an example.

Circuit Generally, a communication path between two points, also referred to as a line, link, or path.

Multiplexing Sharing of a communication path, possibly simultaneously, by several subscribers. Implemented by dividing the path bandwidth into distinct frequency bands (frequency-division multiplexing) or by rotating

a time slot, in turn, among line subscribers (time-division multiplexing).

Network Conglomeration of circuits connected together to form communication paths between subscribers with some common interest.

Protocol Set of rules or conventions governing how data communication takes place over a network, covering a wide range of areas including physical connections, electrical signaling, data representation and packaging, and line access and utilization.

Synchronous Method of synchronizing serial block-level data, characterized by sender and receiver being in synchronization for the duration of the data block. Accomplished by the sender incorporating timing information with the data, allowing the receiver to determine both bit and character times. Contrasts with asynchronous transmission.

THE SUBJECT OF DATA COMMUNICATION in its broadest sense studies the electronic transfer of information between two or more points. Distances between points can range from inches to an interplanetary scale. Here, the interest is in how the points are connected, or networked, and in how network dependencies influence information flow over the network. The technology of data communications touches a large segment of our society. Data systems are an integral part of the business world. Technical people in all disciplines are dependent on computers, and even the average person is likely to have a personal computer at home. All these situations often require the communication of data from one point to another. With this widespread usage in mind, the material presented here is more global in nature, covering neither the deeper technical details of data transmission nor the more theoretical aspects.

I. NETWORK BASICS

A. Network Topologies

Connecting points together is a topological concern and a logical place to start discussing data communication networks. Figure 1 illustrates basic network topologies that may be considered primitive structures, applying to all types of networks. For the time being, a network can be viewed as a simple pathway between points over which information is conveyed. There may or may not be some control element accompanying the network that allows for proper routing of information. The multipoint topology, for instance, has several points connected by a common circuit (communication path or line), with communication being between the controller and points on the network, not between points on the network. In contrast, the bus or ring structures suggest that any point can communicate with any other point on the network. These basic topologies can be linked together to form hierarchical networks of considerable complexity.

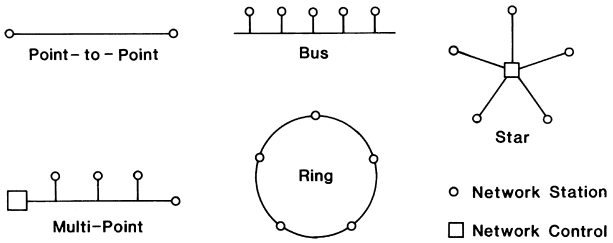


FIGURE 1 Network topologies.

Several circuit configurations can exist within the various network topologies dictating the direction of information flow. A simplex transmission is one where data flows only in one direction. Half-duplex transmission occurs when information can travel in either direction over a circuit, but in only one direction at a time. Full-duplex, or simply duplex, transmission is the simultaneous transfer of information in both directions over a circuit.

B. Network Components

The most basic component of any network is the medium over which information is carried. All networks use some type of conductive media at least somewhere in the data path. However, the conductive path may be broken by a radiative (e.g., microwave) terrestrial or satellite communication link, for example. The basic types of conductive cable are illustrated in Fig. 2. All types except fiber conduct an electrical signal; fiber is used as a "light pipe" to conduct electromagnetic energy in the light spectrum.

For other common network components, Figs. 3 and 4 will serve as a reference for defining both components and some of the associated trade jargon. Starting with the

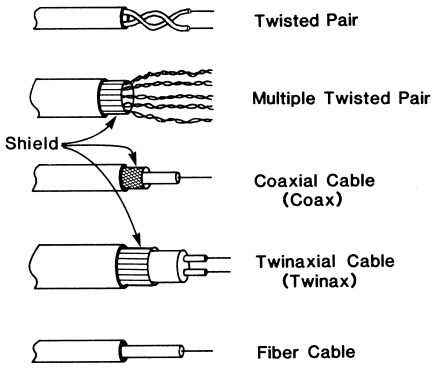


FIGURE 2 Examples of network media.

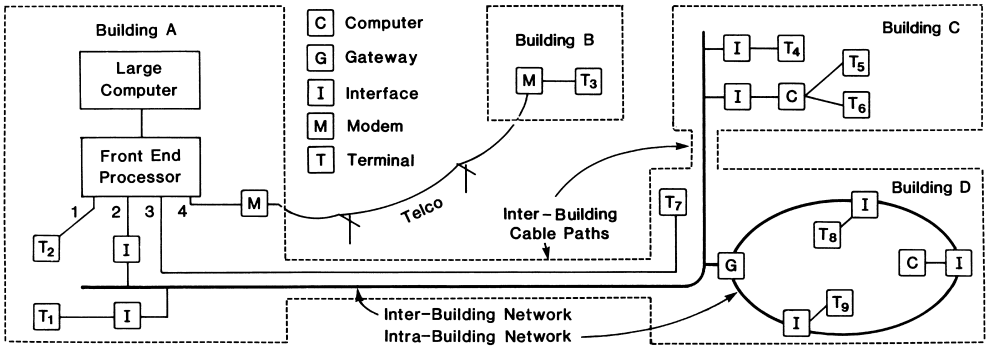


FIGURE 3 Local-area network examples.

local-area environment in Fig. 3, note first the more global aspects. Buildings A, C, and D have a convenient inter-building cable path within which a single bus-type topology network is installed. Building D also happens to have a ring network installed. Building B has no convenient open path to building A and therefore uses a telephone company (telco) circuit to reach the services of the large computer. Telco circuits are a very common type of data connection, since telephone cable is the single most widely installed communication media. Note that this circuit has a modem installed at each end. Loosely, a *modem* is a device that changes the way data is transmitted when it passes from one environment to another. Here, the terminal or front end at either end of the circuit communicates in a digital fashion. However, the telephone circuit conveys data in an analog fashion. The modem allows for the translation from digital to analog and back to digital as data goes from one end of the circuit to the other. Large computers typically have a front-end processor attached, which is really a smaller computer custom designed to handle network traffic into and out of the large computer. Front ends are also called terminal control units.

Note the abundance of interface boxes. Generally speaking, all terminals, computers, or other devices accessing a common network must do so using a common electrical protocol and information format. Interfaces allow this to happen. A terminal, for instance, may not have the proper electrical signal levels to directly connect to a network. An interface will allow for the proper translation of levels. Messages traveling over networks often have other information appended to the message for proper operation. For example, the number of characters in the message or a destination address for the message may have to be added to the message itself. Interfaces can perform these functions also. These functions may be included within a terminal or computer, thus allowing a direct device connection to the network. In any event, proper protocols are enforced somewhere, even if not visible as a separate function.

Other network topologies are evident in Fig. 3. Terminals T_5 and T_6 form a simple star with the computer in building C. A less noticeable star can be seen with terminals T_2, T_3, and T_7 connected to the large computer front end. Connections between T_2, T_3, or T_7 and the large computer, individually, are examples of simple point-to-point circuits. Note the three-level network hierarchy: T_3 in building B could conceivably reach the computer in building D via the front end to the bus network, then through the gateway to the ring network on which the building D computer is connected. One can think of a gateway as an interpreter. The bus and ring networks no doubt have different protocols of use (they speak different languages). The gateway is a device (microcomputer) that allows for a language conversion from one system to another. More formally, this process is called protocol conversion. A similar device to a gateway is the bridge. Linking together two or more distinct networks that use the same protocol is done with a bridge. One common bridge function is to change the data rate of information as it passes from one network to the other. Another function is address translation: a bridge needs to recognize internetwork addresses so that the information can be properly routed from one network to the other. Gateways also accomplish these same functions. Unfortunately, it is generally the rule that differing computers, terminals, and networks do not easily connect to one another, creating a need for a variety of devices that translate protocols, information formats, and electrical characteristics.

The bus or ring networks of Fig. 3 would be suitable candidates for either coaxial or fiber cable. The reason is that for networks having to support a high aggregate data rate (many terminals and/or high-data-rate devices attached), one needs to select a medium that can support that data rate. Both coax and fiber are high-data-rate (bandwidth) cable. Telephone circuits are typically a twisted pair, having a much lower bandwidth. In Fig. 3 applications, one would typically not find radiative links such as microwave because of the short distance. There are

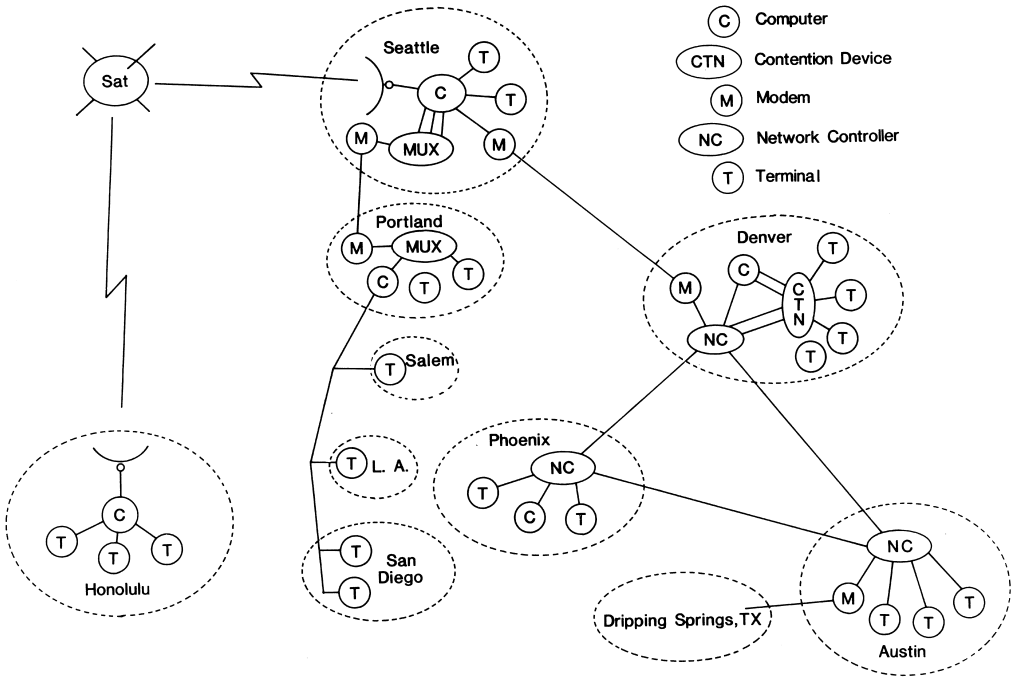


FIGURE 4 Wide-area network examples.

exceptions, of course. For instance, two buildings may need to be connected with a high-bandwidth link, but say a 10-lane freeway exists between them. Then, it may be cheaper to install a radiative-type link using microwave or infrared transmitters and receivers, for example.

Figure 4 changes the environment by increasing network distances. Long-distance data communication is often called "long haul" in trade jargon. "Wide area" is also commonly used. Again, take a global look at the figure. Note that Austin, Denver, and Phoenix are connected in a ring fashion. Point-to-point circuits connect Denver to Seattle and Seattle to Portland. A multipoint (or multidrop) circuit originates at the Portland computer, linking to Salem, Los Angeles, and San Diego. A satellite link connects Seattle to Honolulu. Real intercity data connections are more complex than Fig. 4 suggests. Someone desiring a circuit to another city generally leases one from the telephone company. The circuit leaves the user location often as a simple send/receive pair of wires. Since it looks the same way at the other end of the circuit, one is conditioned to think of long-haul data circuits as simple point-to-point send/receive pairs. In reality, intercity connections are often done by microwave, which is transparent to the circuit subscriber. Satellite links are also microwave, but with one major difference over their terrestrial counterparts. Satellite distances add significant delays to data transmission, which can create problems in certain situations. We will discuss more about this later.

In "short-haul" circuits, where distances are relatively small—i.e., Fig. 3 situations—media cost is often an almost insignificant portion of the overall communication system expense. Multiple connections to a computer, say, may be implemented by simply replicating circuits, one to each terminal. In long-haul situations, circuit costs can become a large part of the communications expense, requiring a different approach. Multiplexing is a technique that allows several devices (terminals, computers) to share one communication path, thus eliminating the need for costly multiple circuits. Figure 4 shows Seattle and Portland connected by one circuit, with each end terminated with a modem and each modem connected to a multiplexer. The modem and multiplexer may be packaged in one unit. The two Portland terminals and computer all share the same circuit via the multiplexer to access a computer system in Seattle. Several forms of multiplexing exist. Frequency-division multiplexing (FDM) takes the allowable bandwidth of a circuit and divides it into several frequency bands. For instance, each Portland terminal could be allocated a unique 300-Hz portion of the circuit bandwidth. One terminal might have 600–900 Hz, another 900–1200 Hz, etc. Time-division multiplexing (TDM) is another form where users have their own unique time slot rather than a frequency slot. View the multiplexer as a switch that sequentially switches to each terminal, say, every millisecond. If there were 10 terminals connected to the Portland TDM, then every 10 msec each terminal

would be allocated 1 msec of time. Note that neither FDM nor TDM techniques discriminate whether or not a terminal is actually using its allotted frequency or time slot. If not in use, that portion of the information spectrum is wasted. An improvement to TDM is to dynamically allocate time slots only to active terminals. A device doing this is known as a statistical time-division multiplexer (STDM), often just called a stat mux, and is a preferred way to multiplex information. By dynamically allocating circuit bandwidth, the number of terminals connected can be increased considerably over earlier techniques, making for much better circuit utilization. There are drawbacks to using a stat mux, even though it efficiently utilizes communication paths. When the number of active users increases, an information bottleneck will occur when the rate of information flow to the multiplexer exceeds the capacity (bandwidth) of the circuit. Buffering is a technique used to manage these peak instances of information flow. Input information passes through a buffer (storage area) within the stat mux on its way to transmission over the circuit. As long as the input rate does not exceed the capacity of the circuit, data flow through the mux is optimal. However, once the input information rate exceeds the line capacity, the buffer begins to fill, queuing the input for later transmission over the link. As the input rate ebbs, the buffer depletes as information is transmitted out of the mux. To visualize the buffering process, compare it to a balloon (buffer) inflating and deflating as its air (information) content varies. If the buffer fills, the mux has no choice but to throttle the input by sending a signal back to the input terminals that it cannot accept any more information. In turn, the terminal must stop its activity, potentially placing an immediate hardship on the user. Obviously, the buffer size, the aggregate input data rate, and the data bandwidth of the connecting circuit must be integrated together for proper operation of the statistical multiplexer.

Contention is another mechanism used to increase utilization of communication resources. In a manner similar to frequency or time slots not being utilized, computer or network ports (entry points to a system) are wasted when users connected to those ports are not using them. A contention device allows several ports to be shared by a larger number of users. Statistically, most of the time not all users are active; therefore, by having a larger number of users contend for a smaller number of ports, everyone will be adequately serviced most of the time, resulting in optimal use of port resources. When the condition arises that a maximum of m ports are being utilized, and the $(m+1)$th user tries to access a system, it simply gets a "busy," i.e., all ports are busy. Figure 4 demonstrates this technique with a contention unit that allows four terminals in Denver to contend for either two ports into a computer or two other ports into the wide-area network. This particular configuration allows for switched contention, where a user can switch between two or more destinations. It is not uncommon for contention to be included within front-end or other equipment.

Finally, a word about switched networks and dial access components in a data communications environment. The occasional computer user often "dials up" the computer service desired using special equipment connected to their telephone circuit. The telephone circuit is part of the telephone company switched network for voice and therefore can reach almost anywhere in the world. The reason dial access is preferred is cost. It is very costly to have a dedicated long-haul circuit, and even more so if it is only sporadically used. A direct-connect modem allows user equipment to directly connect to a telco's switched network by plugging into the same outlet as the standard telephone. An alternate method is via an acoustic coupler into which a telephone handset is inserted. Data are then acoustically coupled via the telephone ear and mouthpiece. In either case, data are converted to audio tones for transmission over the switched network. Figure 4 gives an example. Suppose a user in Dripping Springs, Texas, needs to occasionally access a certain computer in Seattle. To make a more interesting example, further suppose that the Seattle computer does not have a local dial access number but does have remote dial access in Austin via a dial access modem connected to a network controller. The user in Dripping Springs can make a (relatively short distance) call to Austin and ride the network up to Seattle to the desired computer service. This is an example of how a remote computer service can offer "local" access using a communications network. Back in Fig. 3, terminal T_3 could also access building A computer services via a dial connection. The telephone company switched network is commonly used for both local- and wide-area data communication.

II. NETWORK INTELLIGENCE

Up to this point a network has been considered more or less as a convenient pathway to resources such as computers. Not much has been asked of the network; indeed, once installed, it has been viewed as a passive entity. Suppose we wish information such as the amount of data traffic traversing the network or diagnostic data of the network in failure situations. How about asking the network itself to dynamically establish a circuit between a terminal and computer? Or how about asking the network to send a parcel of information to one or more entities on the network? These types of functions considerably elevate the status of a network above simple circuits connecting users to computers. To accomplish this, intelligence must

be somehow added to the network itself. As an example, consider Phoenix, Denver, and Austin linked together in Fig. 4 by network controllers (NCs). Adding the NCs makes the network an intelligent entity in its own right, considerably enhancing the utility of the network. NCs have other names, such as network processors, communication processors, node processors, or node controllers. They all are basically small computers tailored to perform data communication functions for the network. If a link failure between Phoenix and Denver occurs, the network dynamically adjusts, routing all Phoenix/Denver traffic via Austin. A network user would not even necessarily know of such a rerouting. Since all data must pass through the controller, statistics are available, such as the amount of traffic, peak traffic times, outage information, etc. This information allows for more optimal operation of the network.

From a user perspective, the basic function of a network is to direct or switch the user's data to the desired destination. Several switching schemes have evolved as network intelligence has increased. Circuit switches were the earliest, simply establishing point-to-point paths between subscribers. Using a standard telephone amply describes this type of switching, where telephone company equipment establishes a (point-to-point) circuit between two parties. This same concept can apply to data and is often seen in PBX (private branch exchange), where these devices can establish data circuits between subscribers (in addition to their voice functions). Circuit switching establishes dedicated paths. Once done, subscribers need not tell the network where the information is to go; the dedicated path establishes that inherently. An alternative to this scheme is to connect all network subscribers via a common (shared) circuit and to switch the information itself as it flows over the network. This is the basic idea in packet switching. If all network subscribers have an assigned unique address, and if a "packet" of information has an address appended, then information can be delivered across the network between subscribers. A major difference in this switching scheme is that all subscribers time-share the network, but essentially the same end result is accomplished as for circuit switching, i.e., transferring information between points on the network. Obviously, the packet size has a large bearing on the performance of the network. At one extreme, appending N10 bits of address to 10 bits of data results in 50% overhead to get the job done. On the other hand, 10 address bits for 10,000 data bits may reduce overhead significantly, but it might take an extraordinary amount of time to assemble a packet of that size. Selecting optimal packet lengths is an art in itself. Because of the addressing capability, packets can be easily routed around a network to bypass failed elements or congested nodes. Also, an art in this technology is the design of packet routing algorithms. Packets may be statically routed, which essentially means that once a path is established between network subscribers, all packets will take that path for the duration of the session. Adaptive routing algorithms allow packets to be dynamically routed over the network, not necessarily taking the same path as previous packets between the same subscribers. This scheme allows the network to adapt information flow to the current conditions of the network. In Fig. 4, if the Phoenix-to-Denver link started to become congested at some point, part of the packet traffic could be rerouted to Denver via Austin. It is also obvious that adaptive routing can become quite complicated when hundreds of users are shipping packets over a large network. Adaptive packet routing algorithms are often very complex to avoid such subtle things as loosing packets over the network or having packets ride the network forever.

As a final note on items related to network intelligence, message switching can be viewed as a network function. The key concept in message switching is that information does not necessarily have to traverse the network in real time. Up to this point, user connections to resources on a network have been assumed to be responsive, not introducing intolerable delays. Message switching relies on the concept that a message can be stored and forwarded at a later time. Therefore, a storage element has been added, which allows the delayed transmission of the stored message. It should be noted that this function can be provided by a computer resource connected to a network and therefore may not necessarily be an attribute of the network proper.

III. NETWORK AND INFORMATION MANAGEMENT

The basic unit of information in the data world is the bit, having a value of either 0 or 1. Bits of information are transported in two basic ways, either serially or in parallel. Parallel transmission is used primarily in short-distance (a few meters) applications such as between a computer processor and its memory or other peripheral equipment. For example, if 10 paths (wires) exist between 2 points, 10 bits of information can be sent simultaneously (in parallel). In trade for the high information rate allowed by parallel transmission, one pays the added cost of multiple transmission paths. Alternatively, virtually all longer distance data communication (described here) is done serially, i.e., one bit at a time over a single communication path. The details of how 0s and 1s are sent over communication paths is the separate subject of data transmission. For the purposes at hand, all data communication between network points can be viewed as a serial stream of information bits.

This bit stream must be properly packaged and controlled as it is conveyed over the network. A message is a general name for the information package itself.

Message formats depend, to a large extent, on the types of data circuits over which they flow. A point-to-point circuit, for instance, does not require the sender to identify the intended receiver; a multipoint or bus network does. Errors are inevitable in data communication, especially over longer distances. The choice of a particular transmission scheme may depend on how well that scheme can detect errors. The overall flow, control, and synchronization of information over networks are important management functions.

A. Data Synchronization

To transfer information from a sender to a receiver over a network requires that the two entities be synchronized throughout the data transmission time. The most fundamental requirement for synchronization is for both sender and receiver to agree on the transmission rate, i.e., the bits per second (bps). Knowing only the basic bit time, a receiver still does not know when to sample incoming data from the network, i.e., it does not know where an arbitrary bit time starts or ends. Ideally, one would want to sample near the middle of the bit time to stay away from transition levels at the beginning or end. At a higher level, bits are often grouped together to form characters or bytes (1 byte = 8 bits) or, higher yet, blocks or frames of data. A string of characters forming a data block, for instance, would need to be synchronized so that the receiver knows when the block begins. Not knowing this, it is difficult to decompose the block back into the character string. Generally, these basic synchronization problems are solved by transmitting data in one of two standard ways: asynchronously or synchronously.

Asynchronous data transmission is described logically in Fig. 5. The general problem being solved is the synchronization of a character of information between the sender and receiver. Shown is the ASCII character "h" being transmitted over a communication line. In addition to the 7-bit character itself (1101000), a start bit, stop bit, and parity bit are included. Odd parity is shown in the example, which means that an odd number of 1s are in each transmitted character (ensured by proper setting of the parity bit). The parity bit is used by the receiver to detect a limited number of errors that can occur when the character is sent over the communication path. There are variations to the asynchronous format shown, including 8-bit character, no parity, and two rather than one stop bits. Both sender and receiver must agree on a common format in addition to a common bit rate. Once this is done, the receiver constantly looks for a start bit to occur on an idle line. When the start bits occurs, the receiver then has all the information required to determine the character being transmitted over the link. For instance, starting 1.5 bit times after the start bit transition would be an optimal time to sample the first data bit on the line. Since the sender and receiver have agreed on the basic bit time and information format, the receiver then continues to sample succeeding bit times to determine the remaining bits in the character. A stop bit, among other things, ensures that the line goes back to the idle condition, readying it for another character transmission. It is customary in asynchronous data transmission to transmit the least significant bit (LSB) of the character first and the most significant bit (MSB) last. This explains the inversion of the 7-bit character as it traverses the communication line. Note the synchronization overhead for asynchronous transmission. The start, parity, and stop bits are not part of the desired information (character), but are required to effect the transmission. For the example shown, 10 bits are required to send the 7 real information bits, representing a significant overhead, which is constant for every character transmitted. Reiterating, in asynchronous communication, the sender and receiver are only in synchronization between the start and stop bits of

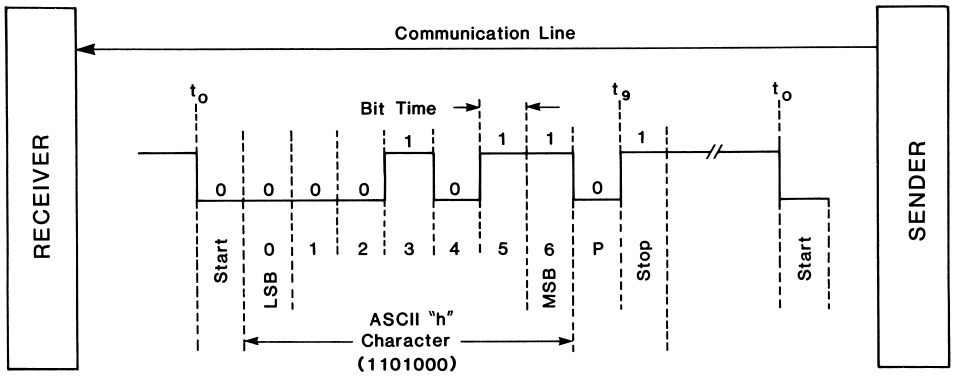


FIGURE 5 Asynchronous data transmission.

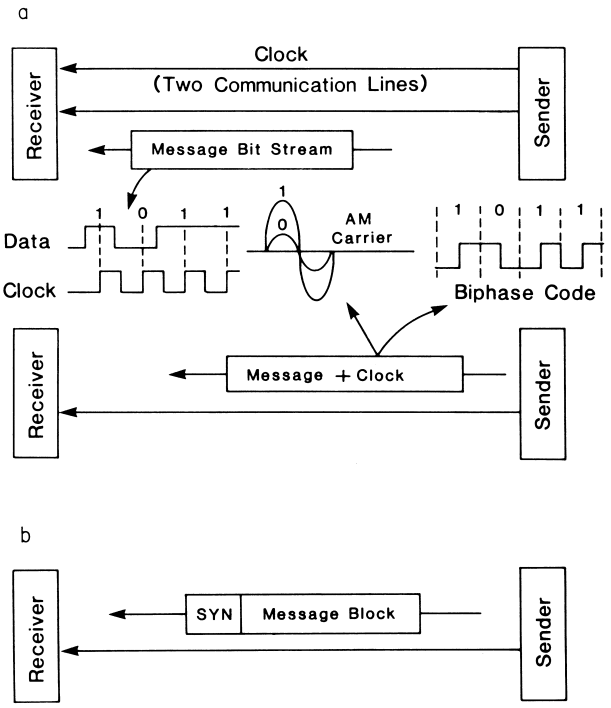


FIGURE 6 Synchronous data transmission: (a) bit synchronization, (b) character synchronization.

each character transmitted. Asynchronous is meant in the sense that the sender and receiver are not in continuous synchronization.

Synchronous data transmission allows larger pieces of information to be synchronized over the network and significantly reduces transportation overhead in comparison to asynchronous transmission. Figure 6a shows three bit-syn-chronization schemes. In short-distance applications, it is common to use both data and clock lines to transport information. The clock line tells the receiver exactly when to sample the data line. In the example shown, positive clock transitions inform the receiver to read out 1 0 1 1 from the data stream. For longer distance applications, data and bit timing information are combind to eliminate the costly need for a separate clocking line. Over certain telco lines, information must be transported in an analog fashion. A rudimentary example is given to demonstrate how bit synchronization can be implemented using an amplitude-modulated (AM) carrier. The carrier amplitude has one of two levels corresponding to the 0 and 1 information bit states. Bit timing is derived from the carrier frequency. In other longer distance applications, data transmission might be done digitally, requiring a different bit-synchronization scheme. One common way is to use a Manchester code, the third example in Fig. 6a. This code is a variation of a class of codes known as biphase codes. Notice that the digital waveform has a low-to-high transition for a logical 1 and a high-to-low transition for a logical 0. Further, there is at least one transition for each bit time; in other words, the code is self-clocking, carrying both clock and data information. Several other self-clocking schemes exist. These examples illustrate how individual bits are synchronized. When bits are logically combined to form characters or bytes as information entities, a requirement is imposed to know where the first bit of a character or byte begins. Figure 6b illustrates one way to accomplish this. The message block is simply a concatenation of information bytes forming a serial bit string eight times the number of bytes long. To achieve byte-level synchronization, a SYN (synchronization) byte is added to the header of the message block. Remember that bit synchronization has already been accomplished. The SYN byte is a unique bit string (for example, 10010110, or 96 hexadecimal) previously agreed on by the sender and receiver. When the bit stream is first detected, the receiver begins a pattern-matching operation to align the incoming bit stream to the known SYN code. When the match occurs, the sender and receiver are then in byte-level synchronization (sync). Once in sync, the succeeding bit stream is easily decomposed into its constituent bytes of information. Synchronization overhead is due to the added SYN byte. As the message block length increases, SYN overhead reduces. Even for short message blocks, overhead is considerably less compared to asynchronous transmission.

B. Synchronous Message Format

The preceding discussion on data synchronization was somewhat idealized to simplify the explanation of basic synchronization concepts. Because synchronous data transmission is so prevalent in longer distance data communication and closely linked to network architecture, message formatting needs to be amplified. Figure 7 illustrates a general synchronous message format, better depicting what actually occurs in practice. At first glance, considerably more overhead is apparent. However, considering that all overhead elements of the message are generally one or two bytes, overhead is still relatively small when the data block is large. Several SYN bytes usually prefix a synchronous message because the network environment can garble the beginning of a message. More than one SYN byte allows the receiver to lock into synchronization even though the first SYN byte may be distorted by

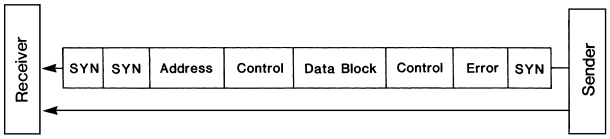


FIGURE 7 Basic synchronous message format.

transmission noise. Over a shared network, address information is necessary to direct the message to the proper destination. The address information of the sender as well might be part of the address. Uses of the control field include defining the message function, signaling the beginning or end of message blocks, acknowledging other transmissions, code switching, status signaling, and issuing commands over the network. The error field is used to detect if a transmitted message has been distorted after traversing the network.

Synchronous formats are distinguished as being either bit or byte oriented. Early formats were byte oriented in design, where management of the message is done at the character level. Character-level management forces code dependency, i.e., the basic network logic depends on the particular code used to represent the characters [e.g., the American standard code for information exchange (ASCII) or the extended binary-coded-decimal interchange code (EBCDIC)]. Subtle problems are also inherent in byte-oriented formats. For example, a data block could contain a control character as a normal part of the sender's data. If not designed to handle this, network control logic could take action at the wrong time in response to the data block control character. Bit-oriented formats treat the message as a bit string, leaving character composition to higher level network functions. This has the immediate advantage of code transparency and is the preferred method in newer format designs.

C. Error Handling

Well-designed electronic equipment is inherently reliable and provides essentially error-free performance. Within itself, the equipment does not generate noise that shows up as performance errors. The vast majority of errors occur when equipments are connected via some media. Further, errors increase as media distances increase. The common television set is a good example. The set itself generally provides good performance for years even though errors are evident almost hourly. Snow, picture distortion, and other errors are not because of the set, but rather, principally, due to environmental noise impacting the communication path. Lightning or human-made electrical signal sources, for example, often inject considerable noise into transmission media.

Because the sole purpose of data communication networks is to provide communication paths between equipments, and since those paths are often very long, error handling is a common and critical function of network design. It is application dependent. In transmitted text, for example, it is easy for a person to detect errors, and for that application, error detection may not be of paramount importance. Errors in numeric data also may be obvious, such as for a singularly misplaced point in an otherwise smooth graph. Errors in financial data can be catastrophic. Error handling is generally discussed in terms of detection and correction, with implementation scheme complexity proportional to the importance of errors being reduced. As a primitive example, a receiving station may echo back each character sent by a terminal operator. Detection and correction, in this case, may rest solely on the operator. Implementing automatic detection is the next step up in complexity, followed by automatic error correction. To add these automatic features, it is necessary to add information to the data being transmitted, which is subsequently used by the receiver to sense and perhaps correct errors introduced over the link. For lower data rates, overhead is generally added to the transmitted data to be used solely for error detection by the receiver, with correction implemented by asking the transmitting station to resend errored data. This simple scheme allows for minimal overhead to the transmitted message and concentrates that overhead at the most critical point, i.e., error detection. Adding more overhead to a message allows the receiving station to both detect and correct certain classes of errors without retransmission. This is known as forward error correction and may be desirable in higher data-rate applications, for instance, where the added overhead can be justified. In practice, adding correction code to a message becomes increasingly more difficult as the code covers more and more error conditions. This is one reason why many strategies place emphasis on the detection code design and use the simple expedient of retransmission for correction. Another reason is the bit error rate performance of typical communication lines. Under normal conditions, 1 bit error in 100,000, or a 1 in 10^5 error rate, is to be expected even for continental distances. The rate can approach 1 in 10^8 as the distance decreases. In many situations, it is inefficient to add correction code overhead to the message when, on the average, that overhead would exceed the overhead in simply resending the message.

The simplest error-handling procedure is the addition of the parity bit, as discussed during asynchronous data transmission and elaborated on in Fig. 8a. This scheme is called a vertical redundancy check (VRC) and has obvious limitations. Any odd number of bit errors can be detected, but any even number of bit errors cannot. Parity is generally added by the sender's equipment just before the character is transmitted and checked by the receiver's equipment upon reception. Typically, parity error detection in the receiving equipment simply flags the error and allows higher level functions to decide what to do about it. In some cases no action is taken.

Longitudinal redundancy checks (LRC) (Fig. 8b) allow blocks of characters to be checked. One way to do this is to form parity over corresponding bits of each character in

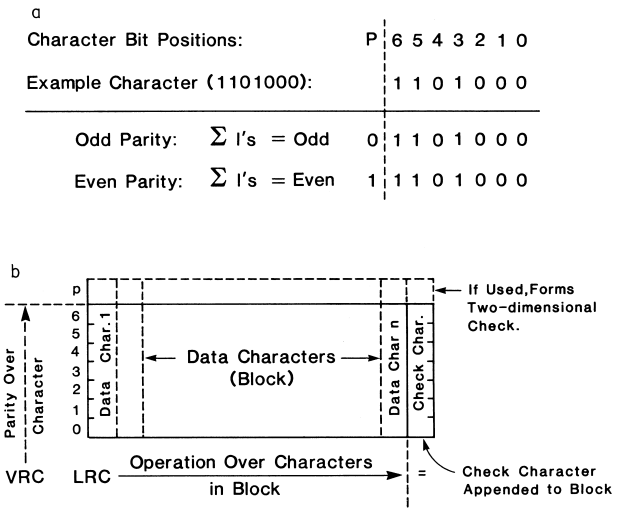


FIGURE 8 Basic error detection schemes: (a) vertical redundancy check over a character, (b) longitudinal redundancy check over a block.

a block of characters. An alternate method is to use some other arithmetic operation. Summing or exclusive-or-ing the character string and using the single character result as a check character is common. The arithmetic check character is appended to the message block for transmission over the network. The receiver performs the same checking operation upon reception of the message block and compares its result to the check character sent by the transmitter. A mismatch signifies an error condition. Note that if each character in the block has an integral parity bit, a two-dimensional check is created with more powerful detection capability.

The error detection schemes considered thus far are commonly used for asynchronous transmission. For synchronous transmission, the cyclic redundancy check (CRC) is common. Errors over typical transmission links tend to occur in bursts. Further, the burst time tends to last over several bit times. CRC techniques are designed to increase the reliability of error detection in this environment. The algorithm to do this is more complex than LRC, taking the user data-bit stream (message block) and dividing it by a predetermined binary number. The quotient forms a check character(s), which is appended to the end of the message block. Figure 7 shows this error detection information as a separate field in the synchronous message format.

D. Access Control

Whenever a communication line is shared among several subscribers, rules need to be established governing line access and utilization. This subject was introduced earlier when multiplexing was described. Dividing the line spectrum into frequency or time slots allows a predetermined number of subscribers to access a common communication line. The multiplexer allows for sharing of one line but effectively provides point-to-point circuits to the individual subscribers. Also note that there is no intersubscriber communication over a multiplexed line. The basic rule to follow in this environment is to maintain a balance between the aggregate subscriber bit rate and the line bandwidth. Described now are techniques to control sharing of communication lines when the full bandwidth of the line is turned over for the sole use of one subscriber at a time in multipoint, bus, or ring network architectures. These environments are more complicated than a point-to-point topology because more rules are required to handle the protocol of turning over use of the line to the various subscribers.

An extension of the TDM concept is to time slot the communication line and allow each subscriber sole use of the line for a predetermined amount of time. The only real difference now is that intersubscriber communication can take place. The slot holder must take care to manage its available time for transmitting one or more messages. The same problem surfaces here as for TDMs, in that time is wasted when a subscriber has nothing to transmit. Workarounds exist to this problem, such as passing a time slot from one subscriber to another. A subscriber not needing its time allotment would then simply pass its time slot to someone else.

Polling or selection (Fig. 9a) was an early development in line control techniques and is still in wide use. In its infancy, data communication was mainframe oriented, where the central computer totally controlled communication line activity in a master–slave relationship with remote stations. Polling is common in wide-area multidrop network topologies. Returning to Fig. 4, as an example, the Portland computer is the master station of the multidrop circuit connecting slave stations in Salem, Los Angeles, and San Diego. The master sequentially polls the slaves to determine if they want to transmit or are ready to receive information to or from the master. The master has a poll list that can be manipulated to control polling priority and timing. Rules of this scheme govern the dialog between, master and slave, a process sometimes called handshaking, which proceeds similar to the following. If the master wants to send data to station C, it sends a poll to C asking if C is ready to receive information. Station C responds with an ACK (positive acknowledgement) if ready, allowing the master to send data. Station C will ACK or NACK (negative acknowledgement) each message for error control. A similar dialog takes place when the master needs to receive information from a slave station. The dialog can become fairly complex to cover the many situations that arise in practice. Polling is widely used and offers

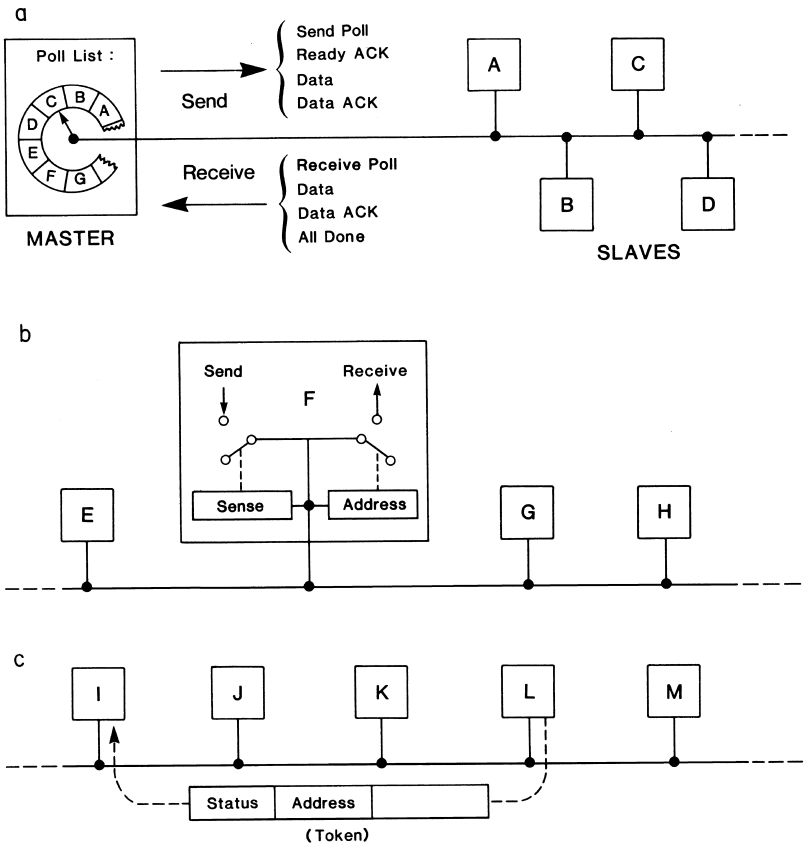


FIGURE 9 Line control techniques: (a) polling, (b) contention, and (c) token passing.

a centralized approach to line management where priorities can be established, since all is under the control of the master. Interstation (slave) communication is generally not a feature of this architecture. Another drawback is the rather high transmission overhead. Even when no actual data transmissions are taking place, polls and NAKs are continually traveling over the network. If time between polls is increased to effect a decrease in overhead, individual subscribers on the network witness increasingly long response times. Thus, both the number of polled stations allowed and the polling rate become factors in designing polled multipoint networks. Redundancy of the master is also a concern, since a failure there literally shuts down the entire polled network.

Consider next the community line in Fig. 9b having the rule that whoever first accesses an idle line gains control of it, i.e., stations are in contention for the line. This is also known as demand access. Station F shows the rough control mechanisms involved with contention line control. A station desiring access continually senses line status and, when idle, switches its operating mode and begins sending data over the link. In addition to sensing, stations also look at the address information of messages traveling over the network. When a station recognizes its own address, it knows to switch its operating mode to receive the message directed to it. Two problems are immediately apparent. First, a rule needs to be established to prevent line hogging. Second, a line going idle may experience several subscribers trying simultaneous access, resulting in collisions. Properly designed, contention is a viable scheme for managing the line even with the collision problem. Collisions are not as frequent as one might think, especially when designs include ways to reduce them, such as having subscribers wait a random amount of time after sensing the line going idle before trying to access it. Subtle problems are inherent in this scheme due to communication line transmission delay. A particular subscriber could gain access to the line and start transmitting its data message. Another subscriber, however, could likewise think it had access simply because transmission delays gave the appearance of an idle line. These timing considerations are inherent in the design of contention networks. Because of delay, contention networks are necessarily limited to campus-level distances. Contention is a fair management approach as every subscriber has equal status and equal opportunity to the line. Fairness, however, is not always a good management approach. Some subscribers might need more frequent access or be allowed longer times

for data transmission to better optimize network and subscriber performance. The inability to assign priorities is a principal disadvantage of the contention approach.

Token passing will be the last line management technique discussed and is generalized in Fig. 9c. This concept was introduced earlier when passing an unused time slot was used to improve efficiency of a time-multiplexed line. A token is a bit string that is passed around the network to the various subscribers. It is an access control mechanism in that whoever receives an idle token has the right to use the network for data transmission. Initially, with no data transfers taking place, the token is idle and declared as idle in its status information field. Destination routing of idle tokens is done either implicitly or explicitly. Stations on a ring topology network inherently pass information from one station to the next in order around the ring. In this case, an idle token does not need destination address information and is said to be passed implicitly. For a bus topology, there is no inherent station ordering, resulting in random access. Here, the token is passed explicitly, or with a destination address included. Note that explicit passing still orders or sequences the stations, resulting in a logical ring over the physical bus network.

A station desiring to transmit information assembles a frame consisting of the data and address information, then waits for an idle token to arrive. When it does, token status is changed to busy, the frame is appended to the token, and the message is sent out over the network. The preceding, although quite general, essentially describes the access protocol involved in token passing networks. Token passing is a natural choice for a ring architecture but can be implemented over a bus as well.

Briefly, access controls can be viewed as operating in a controlled environment or a demand environment. Polling and token passing allow access rights to the network to be passed from one subscriber to another in a controlled, orderly fashion. Collision schemes are demand access situations, where access rights are demanded by each station needing access rather than having the right passed to it by some other station.

E. Flow Control

Up to this point the basic rudiments of networking have been presented. Synchronization, message formatting, error handling, and access control to the communication link are fundamental requirements in data communications. The basic tools now exist to carry on data transfers among subscribers of a common network. Much remains to improve the efficiency of data communication over the network, however. Consider the simple case of a sender that needs to transmit several messages to different subscribers. Considerable time can be wasted if the sender must wait for a positive acknowledgement to each transmitted message before issuing its next message to another station. It can take time for a receiver to check for errors and send back the ACK or NACK. A similar problem exists even if the sender needs to transmit several messages to the same receiver. Implementing controls to handle delayed acknowledgements can improve data flow over the network. Errors further complicate design of the controls. For example, say a receiver correctly reads a message from the network and sends back an ACK. However, due to link noise, suppose the ACK is lost. In this case, the sender will eventually become aware that something is wrong because of the missing acknowledgement and then has no choice but to resend the message that the receiver correctly obtained earlier. Further, if a stream of messages was sent, the ACK block in error must somehow be identified.

From the preceding, it can be seen that flow control is an important function to improved operation of the network. Networks with flow control, at the very least, incorporate sequencing information in the message. Figure 7, for instance, could have an added field identifying the message number. Other enhancements are also possible. The reader is referred to the bibliography for study of these higher level network functions. The topics under discussion in this section generally fall under the heading of data link controls (DLC) in the literature.

F. The ISO Reference Model

Some of the problems and complexity inherent in data communication networking have now been surfaced. There are many ways to implement networking functions, and further, hierarchies of tasks exist as just described for flow control versus lower level functions. Because of this diversity, an array of incompatible networking products evolved in the communication marketplace. Further, additional markets were spawned for gateway products to allow dissimilar networks to intercommunicate. While networking complexity might never allow total standardization, the incompatibility problems faced by consumers made the entire industry keenly aware that some guidance was needed.

In recognition of the basic networking incompatibilities, problems, and lack of direction, the International Standards Organization (ISO) developed their Open Systems Interconnection (OSI) reference model, shown in Fig. 10 (reference: ISO/TC97/SC16/N117 and related documents). The intent of the model is to provide a base level of standardization for the data communications industry in terms of peer groups or common tasks involved when one subscriber talks to another over a network. The OSI model is a hierarchical layered structure with complexity and abstractness increasing from the bottom

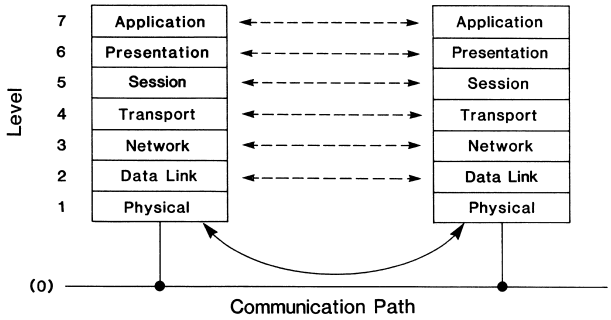


FIGURE 10 ISO reference model.

up. Each layer or level performs a reasonably well-defined network function and is dependent on the layers below to accomplish more elementary functions. The following briefly describes each layer. (The names of some technical specifications are included for further reference.)

Level 0 is simply the physical communication media and is often not listed as a level in itself.

Level 1 governs the physical connection of network entities in regard to basic electrical, mechanical, and signaling protocols. The basic function accomplished is to establish linkage to the physical network and pass logical bits to level 2. Examples of level 1 are EIA-232, EIA-449, and X.21. [EIA specifications are generated by the Electronic Industries Association (EIA) and X standards by the Consultative Committee in International Telegraphy and Telephony (CCITT).]

Level 2 governs much about what the previous section discussed: error handling, polling, and flow control. The idea of level 2 is to provide for reliable and efficient data transfers. Examples are BSC, HDLC, SDLC, ADDCP, and CSMA/CD (see Section IV).

Levels 3–7 cover higher network functions beyond the scope of this article, for which the reader is referred to the bibliography.

The idea of peer communication is now evident from the model. For instance, the data link layers (level 2) of network subscribers communicate with each other in a common, mutually understandable language. This layer does not care about how information was physically passed over the network: that is the job of level 1. Indeed, once a common level language is established, a logical level of communication is formed shown by the dotted lines. Each level can be considered a module with prescribed ways to pass information between adjacent modules. The model is "open" in the sense that the modules are not so tightly specified that they would exclude certain vendor approaches to implementation. In short, the OSI model provides a general architecture for networking.

IV. LOCAL-AREA NETWORKS

As mentioned before, early networks were mainframe oriented. This fostered the master–slave environment where slave stations were tyically dumb terminals. With all the intelligence at the mainframe, none was required at the terminal or in the network itself. Therefore, networking could be accomplished with simple, passive, point-to-point or multipoint circuits. Development of minicomputers started the decentralization of processing power from the mainframe, bringing it closer to individual users. With more intelligence at the user end, the need developed to allow equipments to communicate among themselves, in addition to the mainframe. Network architecture changed accordingly. Networks also had to support increasingly higher data rates to connect sophisticated stations found in the workplace. As the evolution continued, the interconnection of workstations tended to become clustered within a building or small group of buildings to support normal business functions. Local-area networks, or LANs, became the product of this environment.

The concept of LANs was introduced in Fig. 3, which shows the connection of equipment in a local setting. LANs are not rigidly defined, but tend to share the following characteristics:

Data rates are higher, in the range of 1–50 Mbit/sec (M, mega $= 10^6$), to support high-speed communication requirements such as computer-to-computer linkage.
Network geography generally ranges from a building up to a campus-level setting. Distance limitations are primarily a result of the higher data rates.
Network architecture generally allows station-to-station communication in contrast to master–slave relationships.
LANs tend to be privately owned.
Functions other than data communication might utilize the network, such as voice and video communication.

The above characteristics of LANs have a large influence on the media. In the upper megabit range, data transmission must be done over higher quality media such as coaxial cable or fiber. Use of a twisted pair generally requires a reduction in distance or speed, or both. A simplistic view of an LAN is a high-performance cable threaded throughout a building or group of buildings; anyone wishing connection to the LAN runs a cable from their station, tapping into the nearest run of the LAN cable.

Before discussing LANs in more technical detail, it is necessary to first define two general forms of communication which LANs typically employ: baseband or broadband. Baseband refers to dumping the raw digital signal

itself onto the communication line. For instance, a station could have an electrical output of +5 V for a logical 1 and −5 V for a logical 0. In baseband, once the station had access to the line, it could literally switch the line voltage between the +5 and −5 levels to form the electrical analog of the logical bit stream being transmitted. A characteristic of baseband communications is that whoever has access to the line excludes all others. This creates a form of TDM when the line is shared among several subscribers. Broadband communication is similar to cable TV (CATV); indeed, digital broadband is implemented with many CATV industry components. On the TV cable, many channels are broadcast simulataneously by allotting each channel a certain portion of the freqeancy spectrum. To use broadband technology for LANs essentially only requires that a portion of the spectrum be devoted for that use and that a way be provided to convert digital data to a form compatible with analog broadband transmission. To accomplish the latter, a modem is used between LAN subscribers and the broadband cable system. Via the modem, digital data modulates the carrier within that allotted part of the frequency spectrum. Broadband, then, is done by FDM. The principal advantage of broadband is its multiuse capability. It is not uncommon for a broadband network to serve digital LAN, video teleconferencing, TV broadcasts, and other functions. Broadband LANs tend to have a higher number of addressable stations allowed. In practice, however, this feature may not be realized due to the particular protocol in use. For baseband, the principal advantage is cost. If only the LAN function is required, baseband is considerably cheaper at installation and throughout the life of the LAN.

Two of the line protocols already discussed are commonly used for LANs: contention and token passing. Imbedded in these LAN protocols are also many of the other line management techniques discussed earlier, such as error handling and message acknowledgement. Because contention and token passing are so prevalent, they are used here as case examples to further explain higher performance local-area networking. Contention is better known as CSMA/CD in the LAN world, meaning carrier sense multiple access/collision detect. This translates to a network allowing multiple subscribers that sense carrier for access and are able to detect collisions that normally occur with contention. Token passing LANs are either token ring or token bus, corresponding to the two topologies. Both CSMA/CD and token passing are supported by the IEEE 802 standard on local networking, which is another reason for stressing these protocols here. (IEEE 802 is the Institute of Electrical and Electronic Engineer's committee 802 on local networking.)

A. CSMA/CD LANs

Figure 11 shows the approximate implementation of a commercial CSMA/CD LAN using baseband coaxial cable. As an example, a bus topology is shown that allows information to flow equally well in either direction over the bus. Although only two stations are shown for explanation, over 1000 are possible in practice. Each end of the bus is terminated in its characteristic electrical impedance. ISO levels are included to further demonstrate how the model applies to real networks. Levels 3–7 are often called the client layer in CSMA/CD.

All stations are equipped with transceivers that both transmit and receive information over the physical media. The transceivers are responsible for ISO level 1 functions, principally managing the electrical details of getting the bits over the physical media. In this case, the logical bit stream to or from the data link control level would typically be converted to a biphase code, allowing self-clocking of the information over the bus. CSMA/CD bit rates of 10 Mbit/sec are common. Note that the bandwidth of the communication channel (baseband coax) must be capable of sustaining twice this rate, since biphase coding can have up to two transitions per bit time. With the bus topology, all stations examine information passing over the communication path simultaneously (ignoring transmission delay). This feature allows for broadcasting information to one, some, or all stations on the bus.

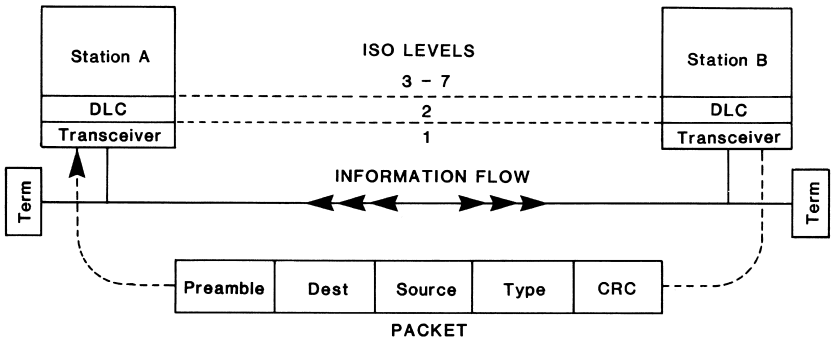


FIGURE 11 CSMA/CD bus baseband LAN.

The information packet begins with a string of bits called the preamble. This is a string of alternating 0s and 1s used to establish bit synchronization of transceivers. Typically, the last two bits of the preamble are 1 1, signifying the end of the preamble and also establishing block synchronization with following address fields. Both destination and source addresses are included in the packet. The type field, explained shortly, is followed by the user data block. The data block is bit oriented and of variable length. The last field, CRC, is the cyclic redundancy checksum used for error detection, as previously explained.

For certain applications, client layer functions may need to intercommunicate over the network. An example might be to establish the protocol being used at the client layer. Passage of this information is accomplished via the type field. To lower levels, the type field looks like part of the data block. Information flow proceeds approximately as the following for transmit and receive functions:

1. Packet assembly begins in the client layer with data, type, and address information, then proceeds to level 2. The DLC calculates the CRC over the packet information at that point, appends it to the packet, and passes on to level 1. The physical layer adds the preamble, performs the biphase code conversion, and sends the packet out over the link.
2. An incoming packet is stripped of the preamble by level 1, with the remaining bit stream passed to the DLC. Using the CRC field, the DLC determines if any transmission errors occurred and checks to see if the packet destination is intended for its particular station. If so, the remaining address, type, and data information are sent on to the client layer.

A shortcoming of CSMA/CD, and one often used to compare against the next discussion of token passing, is the collision problem. Collisions are generally not a problem in lightly loaded networks. As load increases, however, collisions are more frequent, which then begins to impact the throughput performance of the network. CSMA/CD generally includes a feature to help with performance degradation due to collisions, but this does not eliminate it. When the collision rate reaches a certain point, an exponential backoff algorithm is used to increase the time a station waits before trying another transmission. This is a useful approach to smoothing out times when the network becomes quite busy.

B. Token Passing LANs

For comparison to CSMA/CD, a representative commercial token passing LAN is approximated in Fig. 12. A ring topology is shown, which has two distinct differences from a bus. First, information flow is unidirectional over the ring, and second, any station on the ring can cause a break in the communication path. Again, only two stations are shown for simplicity, but many can exist on a single ring in practice. The ring requires the transceiver function of the bus topology to be replaced with individual receivers and drivers to receive and transmit data over the link. This arrangement effectively makes the network a circular linkage of point-to-point circuits. The ISO level functions are basically the same for token and CSMA/CD, needing no further elaboration.

Data transfer over a token ring begins when a station desiring to transmit data receives an idle token. Then the packet status is changed to busy, the information frame is added to the token, and the message is broadcast downlink to succeeding stations. The intended receiving station recognizes its address in the message, resulting in that station capturing the message as it goes by. The message goes around the network until it returns to the original sender. The sender then removes the information frame, changes token status back to idle, and sends the available

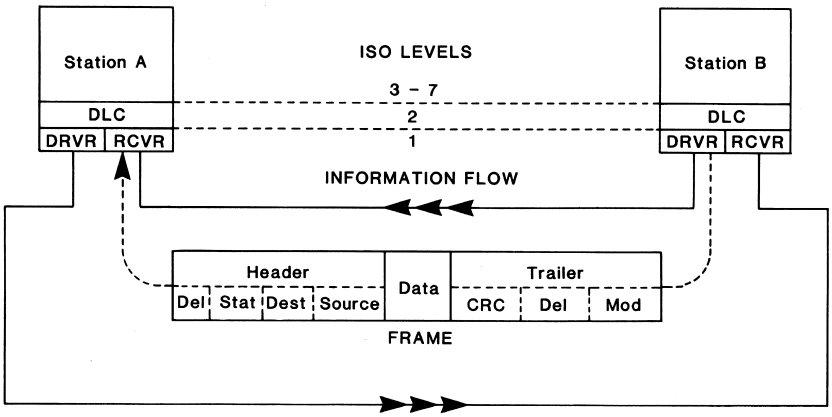


FIGURE 12 Token ring baseband LAN.

token down-link to other stations. The target station that received the message performs the usual tasks of checking for transmission errors and passing the information frame on to higher level functions.

The frame format used in token passing is similar to the packet format of CSMA/CD. The frame is divided into three general fields: header, data block, and trailer. The header and trailer both have delimiters (DEL) that signify the beginning and end of the frame. The remainder of the header includes token status, destination and source addresses, and other control information for link and message management. In conjunction with the trailer delimiter, a modifier (MOD) field terminates the frame and is used for certain control functions, such as error detection and address recognition.

Only the two more popular higher performance LAN approaches have been discussed; both are baseband. In practice, token passing LANs tend to be baseband only, not being implemented over broadband. Both token ring and token bus are in use, both having their respective IEEE 802 standards. CSMA/CD is found in both baseband and broadband implementations.

C. Lower Performance LANs

Several LAN strategies exist in the range of under 1 Mbit/sec. With lower data rates, twisted-pair transmission medium comes back into play because of its low cost and widespread use. As a case example for lower performance LAN technology, digital PBX-type technology is chosen here for several reasons. First, these devices are immediately usable for digital data transport. If an older analog-type switch exists, it can be easily replaced with a digital version using the same cable plant. Second, the existing cable plant of the PBX generally reaches everyone in the local community to service their telephone needs and, as a result, is a significant resource when evaluating LAN needs. For example, if 1000 stations needed connection over an LAN, considerable time and expense would be devoted just to install cable. If the capabilities of the PBX and cable plant match the LAN requirements, no additional wiring may be necessary. As LAN data rates decrease, LAN functions begin to merge with data applications running over telephone circuits. It then becomes difficult to define the boundaries of an LAN. It becomes more correct to speak of local networking in general rather than about an "LAN." Finally, this environment also serves to introduce wide-area networking, which immediately follows this topic.

Referring to Fig. 13, a local setting is shown about a telephone company central office. All telephone-subscriber voice and data lines for a community are connected into a central office. These lines are referred to as local loops, being simple twisted-pair circuits. At the central office, switching equipment exists that connects a subscriber to other local or distant subscribers as directed by the caller's

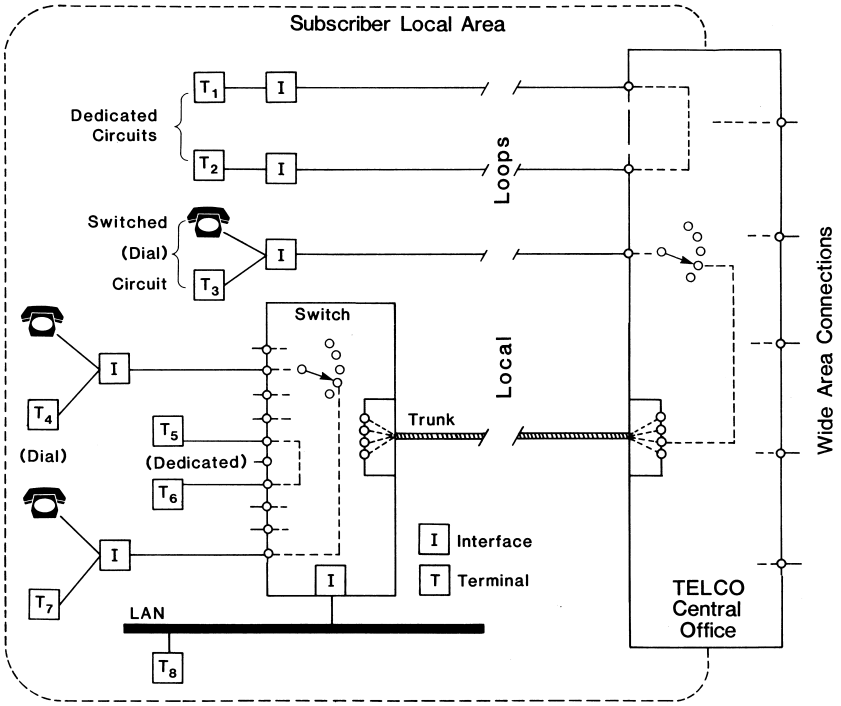


FIGURE 13 Private exchanges, local loops.

dialed digit sequence (telephone number). Lines to the central office can functionally be classified into switched (dial) or dedicated circuits. Both are physically the same type of twisted-pair cable. A switched or dial circuit is one that is tied into the switching equipment at the central office—i.e., it relies on the caller to dial the desired connection. A dedicated circuit is one that is permanently connected: source and destination circuits at the central office are hardwired together, dedicating the path to the subscribers involved. The PBX is a smaller version of the central office switching equipment and is commonly used in many business applications. It allows all telephones within a building, or community of buildings, to be interconnected. Geographically, this situation is precisely the same as that for an LAN. PBXs are linked to central office functions via trunk lines. Figure 13 shows a generic switch that could be one of several things, one being the PBX. At a more sophisticated level, it could be a PNX, or private network exchange, which may incorporate higher level LAN functions. Computer base exchange, or CBX, is also a term used to denote a more sophisticated switch. An interface is shown from the switch to a higher performance LAN to illustrate a higher level function than a simple PBX.

Several examples are given in Fig. 13 to illustrate lower performance local networking. Terminals T_1 and T_2 are connected as a point-to-point dedicated circuit via a hardwired connection at the central office. Either terminal could just as easily have been connected to some remote city via the telephone company wide-area circuitry. Terminal T_3 is connected to an interface along with a standard telephone. The user can dial another subscriber to establish a connection, then switch over to the data terminal. There are several variations of this capability as discussed earlier. Terminal T_4 is like terminal T_3, but connected to a local switch rather than directly to the central office. Terminals T_5 and T_6 are likewise similar to terminals T_1 and T_2, being connected via the local switch. With this brief overview, it is easy to visualize many ways to make local data connections using switch devices.

V. WIDE-AREA NETWORKING

Figure 4 introduced wide-area networking in general, ignoring some of the detail involved in establishing a data path. Figure 14 takes up where Fig. 13 left off, in terms of wide-area connections, and adds enough detail to visualize how long-haul data circuits are implemented. The more common long-haul circuits use common carrier services of the telephone company. This is an old term originating from the idea of providing a carrier service common to the population at large. Originally, telephone service was classed as a basic carrier, as were ship and rail services. Later, data services were offered over the same analog transmission services, followed later yet with digital data services. Shown in Fig. 14 are several general ways a terminal in city A might be connected to a computer in city B. If a dial connection is desired, the caller dials the destination telephone number, which results in a path being established through all the switching toll centers between the two cities. A dedicated circuit can be ordered from the carrier, established by hardwiring connections along the path between subscribers. As the distance increases between points, it is more likely that connections include high-bandwidth transmission media, rather than simple switched or dedicated copper wire connections. Terrestrial microwave is an example, providing high-bandwidth

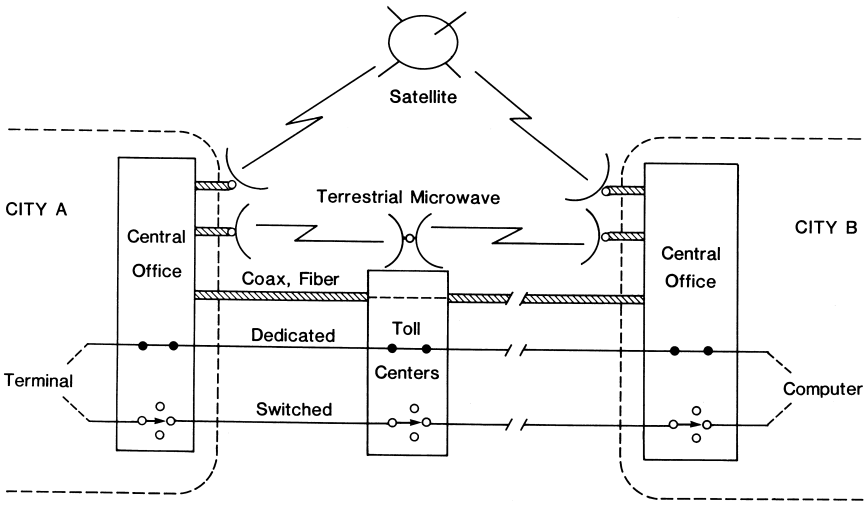


FIGURE 14 Wide-area networking.

communication for voice, data, and video between cities. Microwave dishes are a common sight on top of buildings and high-terrain points, with about 30 miles between dishes being the maximum. Another example is coax or fiber cable, typically buried between cities. Information traffic between cities, then, is typically multiplexed over the high-bandwidth capabilities of microwave or cable. Finally, long-haul circuits can be implemented by multiplexing a subscriber over a high-bandwidth communication satellite link.

Each of the several ways to implement wide-area circuits have their pros and cons. Switched circuits tend to be noisy and are used in sporadic-need, lower data-rate applications. Dedicated terrestrial links are of higher quality with less noise and are used in applications where high time connectivity and low circuit delays are important. Satellite links are generally implemented via geosynchronous satellites, which are in stationary orbit relative to points on the earth, about 22,300 miles in altitude. This distance requires considerable time for signal propagation, averaging about 540 msec and peaking to 900 msec in some situations. To visualize the impact of this, consider a polled circuit implemented via a satellite circuit. Short blocks of information can require only a few milliseconds of time, small compared to satellite link transit time. In a polled environment, considerable time would be wasted waiting on satellite link delays.

With a variety of communication services available, subscribers can order data circuits with a range of data rates and performances. Using commonly available equipment and existing subscriber cable services, it is easy to implement circuits with data rates of 50 kbit/sec (k, kilo $= 10^3$). With a little more effort, over a megabit per second can be provided for long-haul applications.

A. Analog and Digital Facilities

Originally, data sent over wide-area links had to be converted to analog form to be compatible with common carrier equipment. Referring back to Fig. 13, the interfaces shown would be modems that modulate analog carrier with data to ship information over the network. As technology improved, common carriers implemented digital data services that allowed digital data to be directly interfaced to the carriers' equipment. A digital interface is required to multiplex the subscriber data into the carrier's bit stream, but no digital/analog conversion takes place. This reduces cost, as digital multiplexing is cheaper than modulating and demodulating data, and also benefits from the higher reliability inherent in digital equipment.

To understand some of the rationale of bit rates used in digital data transmission, it is necessary to discuss digitizing voice. Telephone voice circuits are generally analog, at least up to the central office. Past there, analog signals may be digitized so that they can be transmitted over the telephone company's digital facilities. To transfer voice over telephone circuits, a bandwidth of about 4 kHz is needed to carry a reasonable representation of the talker's voice characteristics. To digitize the voice occurring over the line requires that it be sampled at twice the highest frequency, or 8000 times per second. At each sample time, the signal amplitude must be converted to a binary number. Seven bits are used for the conversion, which allows discrimination of 128 different amplitude levels. Another bit is added for control, bringing the total to 8 bits for each sampled point. Twenty-four voice channels are customarily time multiplexed (TDM) for transmission over digital facilities. The multiplexing takes place by sampling each of the 24 channels, in order, producing 192 bits of information. Another bit is added for control, bringing the total to 193 bits. Since this needs to be done 8000 times per second, a transmission rate of 1.544 Mbit/sec results. This is called a T1 carrier. Rough multiples of the T1 service are used to form a hierarchy of digital transmission, as shown in Fig. 15. Data rates increase as more subscribers are multiplexed together. A hierarchy of media also results, starting with the basic twisted pair at the subscriber end up to coax or fiber, which is required to carry the higher data rates. Basic synchronous digital data service is offered by carriers at data rates of 2.4, 4.8, 9.6, and 56 kbit/sec over standard local loops to the central office. It is also possible to access the T1 carrier for high-data-rate applications.

B. Integrated-Services Digital Network

The cost effectiveness, functionality, and inherent reliability of digital transmission result in a continued migration of digital connectivity to end-user applications. Direct digital interfaces to end-user equipment are becoming available from the carriers for a widening array of applications. A common example is the digital telephone. This type of phone, within itself, converts the voice analog signal to a digital form, allowing the device to directly connect to a digital transmission channel. As an example of higher complexity, video teleconferencing is done by bringing a digital "pipe" to a subscriber's premises of sufficient bandwidth to allow the transmission of digitized video signals. The concept of directly connecting a variety of end-user devices to digital carrier services is known as integrated-services digital network, or ISDN.

C. Other Wide-Area Services

Up to this point, only common carrier services have been discussed. Specialized common carriers (SCCs) also exist, offering only specialized services, not the wider offerings

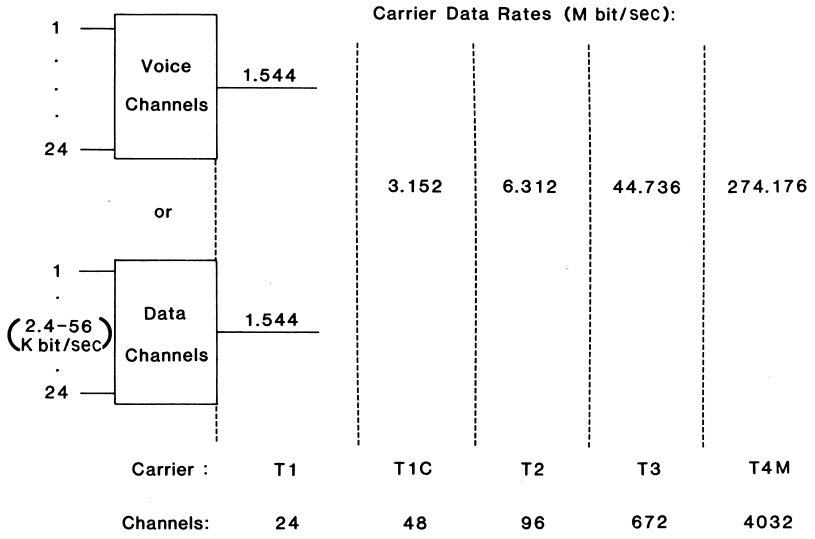


FIGURE 15 Digital data services.

of the common carriers. For example, an independently owned, high-bandwidth terrestrial microwave could be installed with a tie-point to central office equipment (Fig. 14) leased from a common carrier. This would allow the specialized carrier (independent microwave) access to local loop service to provide its specialized services to the general population, which keeps it in the class of common carrier, but specialized to some particular application.

Another example of carrier services is the value-added carrier (VAC). As an example, someone owning a computer could implement a special computer database containing certain information that might be of interest to a wide variety of people. One way to provide access to the information would be to install dial modems to the computer, allowing anyone with a telephone to tie into the database. Passing an account number, say, after dialing in would allow the provider to bill for services rendered. One problem with this approach is that many potential customers might be turned away by the long-distance telephone charges. If the service is popular over a large geographical area, even globally, another approach would be to distribute the access mechanism. The service could lease long-haul circuits from the common carriers and install dial or leased line access equipment over a wide area, providing local access to residents in the cities serviced. In other words, the common carriers have been used to implement a service that added a value to the basic common carrier service.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICIAL NEURAL NETWORKS • BASEBAND DIGITAL LOOPS • COMPUTER NETWORKS • DATA TRANSMISSION MEDIA • INFORMATION THEORY • MICROWAVE COMMUNICATIONS • MULTIPLEXING • SIGNAL PROCESSING, GENERAL • TELECOMMUNICATIONS • TOPOLOGY, GENERAL

BIBLIOGRAPHY

Black, U. D. (1983). "Data Communications, Networks and Distributed Processing," Reston Publishing, Reston, VA.

Dixon, R. C., Strole, N. C., and Markov, J. D. (1983). "A token-ring network for local data communications," *IBM Syst. J.* **22**(1/2), 47–62.

Kreager, P. S. (1983). "Practical Aspects of Data Communication," McGraw-Hill, New York.

Stallings, W. (1985). "Data and Computer Communications," MacMillan Co, New York.

Tanenbaum, A. S. (1981). "Computer Networks," Prentice-Hall, Englewood Cliffs, NJ.

IBM (March 1984). "A Building Planning Guide for Communication Wiring," G320-8059-1, IBM Publication.



Optical Fiber Communications

B. K. Tariyal
AT&T Technologies, Inc.

A. H. Cherin
AT&T Bell Laboratories

I. Introduction
II. Classification of Optical Fibers and Attractive Features
III. Fiber Transmission Characteristics
IV. Optical Fiber Cable Manufacturing
V. Sources and Detectors
VI. Source Coupling, Splices, and Connectors
VII. Applications of Optical Fiber Communications
VIII. Recent Developments

GLOSSARY

Attenuation Decrease of average optical power as light travels along the length of an optical fiber.

Bandwidth Measure of the information-carrying capacity of the fiber. The greater the bandwidth, the greater the information-carrying capacity.

Barrier layer Layer of deposited glass adjacent to the inner tube surface to create a barrier against OH diffusion.

Chemical vapor deposition Process in which products of a heterogeneous gas–liquid or gas–solid reaction are deposited on the surface of a substrate.

Cladding Low-refractive-index material that surrounds the fiber core.

Core Central portion of a fiber through which light is transmitted.

Cutoff wavelength Wavelength greater than which a particular mode ceases to be a bound mode.

Dispersion Cause of distortion of the signal due to different propagation characteristics of different modes, leading to bandwidth limitations.

Graded index profile Any refractive index profile that varies with radius in the core.

Microbending Sharp curvatures involving local fiber axis displacements of few micrometers and spatial wavelengths of a few millimeters. Microbending causes significant losses.

Mode Permitted electromagnetic field pattern within an optical fiber.

Numerical aperture Acceptance angle of the fiber.

Optical repeater Optoelectric device that receives a signal and amplifies it and retransmits it. In digital systems the signal is regenerated.

COMMUNICATIONS using light as a signal carrier and optical fibers as transmission media are termed optical fiber communications. The applications of optical fiber communications have increased at a rapid rate since the

first commercial installation of a fiber-optic system in 1977. Today every major telecommunication company is spending millions of dollars on optical fiber communication systems. In an optical fiber communication system, voice, video, or data are converted to a coded pulse stream of light using a suitable light source. This pulse stream is carried by optical fibers to a regenerating or receiving station. At the final receiving station the light pulses are converted to electric signals, decoded, and converted into the form of the original information. Optical fiber communications are currently used for telecommunications, data communications, military applications, industrial controls, medical applications, and cable television.

I. INTRODUCTION

Since ancient times, humans have used light as a vehicle to carry information. Lanterns on ships and smoke signals or flashing mirrors on land are early examples of how humans used light to communicate. It was just over 100 years ago that Alexander Graham Bell (1880) transmitted a telephone signal a distance greater than 200 m using light as the signal carrier. Bell called his invention a "Photophone" and obtained a patent for it. Bell, however, wisely gave up the photophone in favor of the electric telephone. Photophone at the time of its invention could not be exploited commercially because of two basic drawbacks: the lack of a reliable light source and the lack of a dependable transmission medium.

The invention of the laser in 1960 gave a new impetus to the idea of lightwave communications (as scientists realized the potential of the dazzling information-carrying capacity of these lasers). Much research was undertaken by different laboratories around the world during the early 1960s on optical devices and transmission media. The transmission media, however, remained the main problem, until K. C. Kao and G. A. Hockham in 1966 proposed that glass fibers with sufficiently high-purity core surrounded by a lower-refractive-index cladding could be used for transmitting light over long distances. At the time, available glasses had losses of several thousand decibels per kilometer. In 1970, Robert Maurer of Corning Glass Works was able to produce a fiber with a loss of 20 dB/km. Tremendous progress in the production of low-loss optical fibers has been made since then in the various laboratories in the United States, Japan, and Europe, and today optical fiber communication is one of the fastest growing industries. Optical fiber communication is being used to transmit voice, video, and data over long distance as well as within a local network.

Fiber optics has emerged as the method of choice for many communications applications. The biggest advantage of a lightwave system is its tremendous information-carrying capacity. There are already systems that can carry several thousand simultaneous conversations over a pair of optical fibers thinner than human hair. In addition to this extremely high capacity, the lightguide cables are lightweight, immune to electromagnetic interference, and potentially very inexpensive.

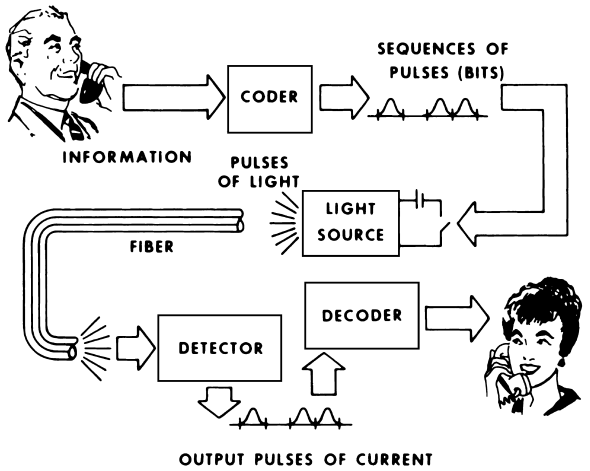


FIGURE 1 Schematic diagram of a lightwave communications systems.

A lightwave communication system (Fig. 1) consists of a transmitter, a transmission medium, and a receiver. The transmitter takes the coded electronic signal (voice, video, or data) and converts it to the light signal, which is then carried by the transmission medium (an optical fiber cable) to either a repeater or the receiver. At the receiving end the signal is detected, converted to electrical pulses, and decoded to the proper output. This article provides a brief overview of the different components used in an optical fiber system, along with examples of various applications of optical fiber systems.

II. CLASSIFICATION OF OPTICAL FIBERS AND ATTRACTIVE FEATURES

Fibers that are used for optical communication are waveguides made of transparent dielectrics whose function is to guide light over long distances. An optical fiber consists of an inner cylinder of glass, called the core, surrounded by a cylindrical shell of glass of lower refractive index, called the cladding. Optical fibers (lightguides) may be classified in tems of the refractive index profile of the core and whether one mode (single-mode fiber) or many modes (multimode fiber) are propagating in the guide (Fig. 2). If the core, which is typically made of a high-silica-content glass or a multicomponent glass, has a uniform refractive

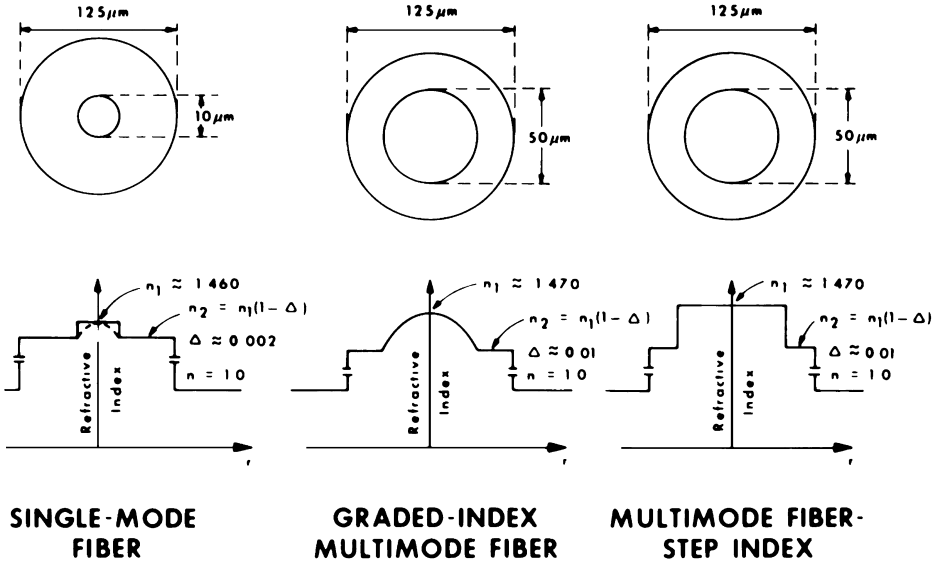


FIGURE 2 Geometry of single-mode and multimode fibers.

index, n_1, it is called a *step-index fiber*. If the core has a nonuniform refractive index that gradually decreases from the center toward the core–cladding interface, the fiber is called a *graded-index fiber*. The cladding surrounding the core has a uniform refractive index, n_2, that is slightly lower than the refractive index of the core region. The cladding of the fiber is made of a high-silica-content glass or a multicomponent glass. Figure 2 shows the dimensions and refractive indices for commonly used telecommunication fibers. Figure 3 enumerates some of the advantages, constraints, and applications of the different types of fibers. In general, when the transmission medium must have a very high bandwidth—for example, in an undersea or long-distance terrestrial system—a single mode fiber is used. For intermediate system bandwidth requirements between 200 MHz/km and 2 GHz/km, such as found in intracity trunks between telephone central offices or in local area networks, either a single-mode or graded-index multimode fiber would be the choice. For applications such as short data links where lower bandwidth requirements are placed on the transmission medium, either a graded-index or a step-index multimode fiber may be used.

Because of their low loss and wide bandwidth capabilities, optical fibers have the potential for being used wherever twisted wire pairs or coaxial cables are used as the transmission medium in a communication system. If an engineer were interested in choosing a transmission medium for a given transmission objective, he or she would tabulate the required and desired features of alternate technologies that may be available for use in the

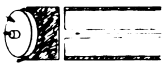
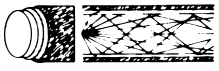
	SINGLE-MODE FIBER	GRADED-INDEX MULTIMODE FIBER	STEP-INDEX MULTIMODE FIBER
Cladding / Core / Protective Plastic Coating			
SOURCE	LASER PREFERRED	LASER or LED	LASER or LED
BANDWIDTH	VERY VERY LARGE > 3 GHz km	VERY LARGE 200 MHz 3 GHz km	LARGE < 200 MHz km
SPLICING	DIFFICULT DUE TO SMALL CORE	DIFFICULT BUT DOABLE	DIFFICULT BUT DOABLE
EXAMPLE OF APPLICATION	SUBMARINE CABLE SYSTEM	TELEPHONE LOOP DISTRIBUTION SYSTEM	DATA LINKS
COST	LEAST EXPENSIVE	MOST EXPENSIVE	EXPENSIVE

FIGURE 3 Applications and characteristics of fiber types.

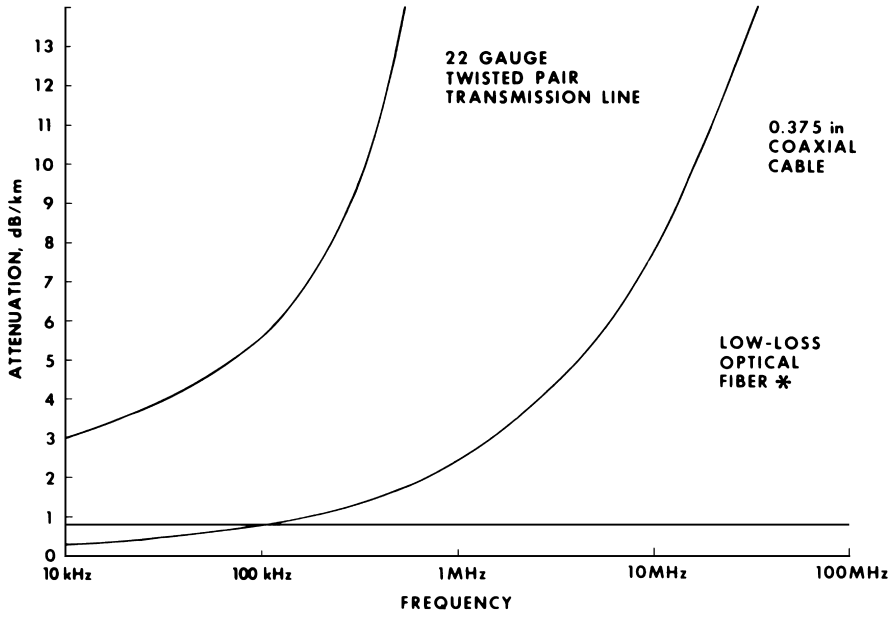


FIGURE 4 Attenuation versus frequency for three different transmission media. Asterisk indicates fiber loss at a carrier wavelength of 1.3 μm.

applications. With that process in mind, a summary of the attractive features and the advantages of optical fiber transmission will be given. Some of these advantages include (a) low loss and high bandwidth; (b) small size and bending radius; (c) nonconductive, nonradiative, and noninductive; (d) light weight; and (e) providing natural growth capability.

To appreciate the low loss and wide bandwidth capabilities of optical fibers, consider the curves of signal attenuation versus frequency for three different transmission media shown in Fig. 4. Optical fibers have a "flat" transfer function well beyond 100 MHz. When compared with wire pairs of coaxial cables, optical fibers have far less loss for signal frequencies above a few megahertz. This is an important characteristic that strongly influences system economics, since it allows the system designer to increase the distance between regenerators (amplifiers) in a communication system.

The small size, small bending radius (a few centimeters), and light weight of optical fibers and cables are very important where space is at a premium, such as in aircraft, on ships, and in crowded ducts under city streets.

Because optical fibers are dielectric waveguides, they avoid many problems such as radiative interference, ground loops, and, when installed in a cable without metal, lightning-induced damage that exists in other transmission media.

Finally, the engineer using optical fibers has a great deal of flexibility. He or she can install an optical fiber cable and use it initially in a low-capacity (low-bit-rate) system. As the system grows, the engineer can take advantage of the broadband capabilities of optical fibers and convert to a high-capacity (high-bit-rate) system by simply changing the terminal electronics. A comparison of the growth capability of different transmission media is shown in Table I. For the three digital transmission rates considered (1.544, 6.312, and 44.7 Mbit/sec), the loss of the optical fiber is constant. The loss of the metallic transmission lines, however, increases with increasing transmission rates, thus limiting their use at the higher bit rates. The optical fiber system, on the other hand, could be used at all bit rates and can grow naturally to satisfy system needs.

III. FIBER TRANSMISSION CHARACTERISTICS

The proper design and operation of an optical communication system using optical fibers as the transmission medium requires a knowledge of the transmission characteristics of the optical sources, fibers, and interconnection devices (connectors, couplers, and splices) used to join lengths of fibers together. The transmission criteria that affect the choice of the fiber type used in a system are signal attenuation, information transmission capacity (bandwidth), and source coupling and interconnection efficiency. Signal attenuation is due to a number of loss mechanisms within the fiber, as shown in Table II, and due to the losses occurring in splices and connectors. The

TABLE I Growth Capability, Transmission Media Comparisons

Transmission media	Loss (in dB/km) at half-bit-rate frequency (digital transmission rates)		
	T1 (1.544 Mbit/sec)	T2 (6.312 Mbit/sec)	T3 (44.736 Mbit/sec)
26-Gauge twisted wire pair	24	48	128
19-Gauge twisted wire pair	10.8	21	56
0.375-in.-Diameter coaxial cable	2.1	4.5	11
Low-loss optical fiber[a]	0.75	0.75	0.75
Single mode	0.35	0.35	0.35

[a] Fiber loss at a carrier wavelength of 1.3 μm.

information transmission capacity of a fiber depends on dispersion, a phenomenon that causes light that is originally concentrated into a short pulse to spread out into a broader pulse as it travels along an optical fiber. Source and interconnection efficiency depend on the fiber's core diameter and its numerical aperture, a measure of the angle over which light is accepted in the fiber.

Absorption and scattering of light traveling through a fiber leads to signal attenuation, the rate of which is measured in decibels per kilometer (dB/km). As Fig. 5 shows, for both multimode and single-mode fibers, attenuation depends strongly on wavelength. The decrease in scattering losses with increasing wavelength is offset by an increase in material absorption such that attenuation is lowest near 1.55 μm (1550 nm).

The measured values given in Table III are probably close to the lower bounds for the attenuation of optical fibers. In addition to intrinsic fiber losses, extrinsic loss mechanisms, such as absorption due to impurity ions, and microbending loss due to jacketing and cabling can add loss to a fiber.

The bandwidth or information-carrying capacity of a fiber is inversely related to its total dispersion. The total dispersion in a fiber is a combination of three components: intermodal dispersion (modal distortion), material dispersion, and waveguide dispersion.

Intermodal dispersion occurs in multimode fibers because rays associated with different modes travel different effective distances through the optical fiber. This causes light in the different modes to spread out temporally as it travels along the fiber. Modal delay distortion can severely limit the bandwidth of a step-index multimode fiber to the order of 20 MHz/km. To reduce modal delay distortion in multimode fibers, the core is carefully doped to create a graded (approximately parabolic shaped) refractive index profile. By carefully designing this index profile, the group velocities of the propagating modes are nearly equalized. Bandwidths of 1.0 GHz/km are readily attainable in commercially available graded index multimode fibers.

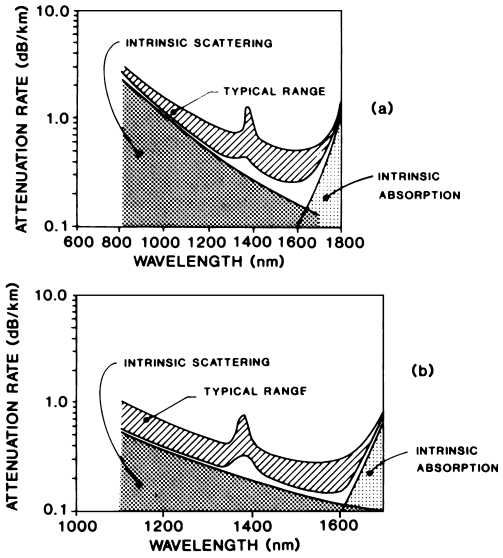


FIGURE 5 Spectral attenuation rate. (a) Graded-index multimode fibers. (b) Single-mode fibers.

TABLE II Loss Mechanisms

Intrinsic material absorption loss
Ultraviolet absorption tail
Infrared absorption tail
Absorption loss due to impurity ions
Rayleigh scattering loss
Waveguide scattering loss
Microbending loss

TABLE III Best Attenuation Results (in dB/km) in Ge–P–SiO_2-Core Fibers

Wavelength (nm)	$\Delta \approx 0.2\%$ (Single-mode fibers)	$\Delta \approx 1.0\%$ (Graded-index multimode fibers)
850	2.1	2.20
1300	0.27	0.44
1500	0.16	0.23

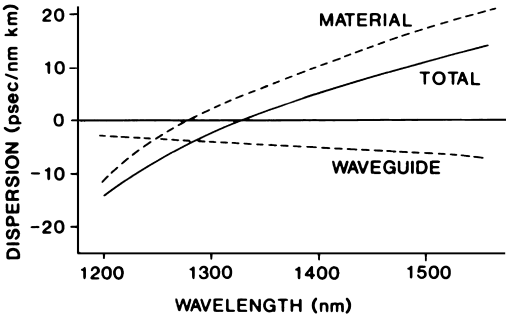


FIGURE 6 Single-mode step-index dispersion curve.

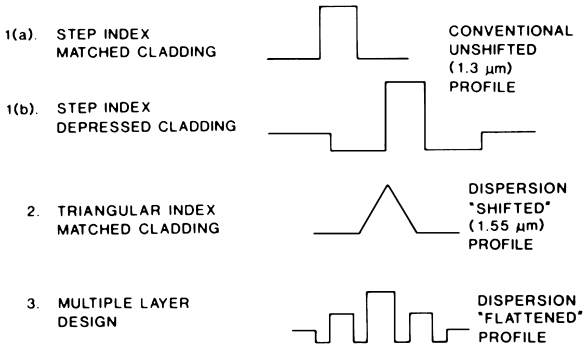


FIGURE 7 Single-mode refractive-index profiles.

The most effective way to eliminate intermodal dispersion is to use a single-mode fiber. Since only one mode propagates in a single mode fiber, modal delay distortion between modes does not exist and very high bandwidths are possible. The bandwidth of a single-mode fiber, as mentioned above, is limited by the combination of material and waveguide dispersion. As shown in Fig. 6, both material and waveguide dispersion are dependent on wavelength.

Material dispersion is caused by the variation of the refractive index of the glass with wavelength and the spectral width of the system source. Waveguide dispersion occurs because light travels in both the core and cladding of a single-mode fiber at an effective velocity between that of the core and cladding materials. The waveguide dispersion arises because the effective velocity, the waveguide dispersion, changes with wavelength. The amount of waveguide dispersion depends on the design of the waveguide structure as well as on the fiber materials. Both material and waveguide dispersion are measured in picoseconds (of pulse spreading) per nanometer (of source spectral width) per kilometer (of fiber length), reflecting both the increases in magnitude in source linewidth and the increase in dispersion with fiber length.

Material and waveguide dispersion can have different signs and effectively cancel each other's dispersive effect on the total dispersion in a single-mode fiber. In conventional germanium-doped silica fibers, the "zero-dispersion" wavelength at which the waveguide and material dispersion effects cancel each other out, occurs near 1.30 μm. The zero-dispersion wavelength can be shifted to 1.55 μm, or the low-dispersion characteristics of a fiber can be broadened by modifying the refractive-index profile shape of a single-mode fiber. This profile shape modification alters the waveguide dispersion characteristics of the fiber and changes the wavelength region in which waveguide and material dispersion effects cancel each other. Figure 7 illustrates the profile shapes of "conventional," "dispersion-shifted," and "dispersion-flattened" single-mode fibers. Single-mode fibers operating in their zero-dispersion region with system sources of finite spectral width do not have infinite bandwidth but have bandwidths that are high enough to satisfy all current high-capacity system requirements.

IV. OPTICAL FIBER CABLE MANUFACTURING

Optical fiber cables should have low loss and high bandwidth and should maintain these characteristics while in service in extreme environments. In addition, they should be strong enough to survive the stresses encountered during manufacture, installation, and service in a hostile environment. The manufacturing process used to fabricate the optical fiber cables can be divided into four steps: preform fabrication, fiber drawing and coating, fiber measurement, and fiber packaging.

A. Preform Fabrication

The first step in the fabrication of optical fiber is the creation of a glass preform. A preform is a large blank of glass several millimeters in diameter and several centimeters in length. The preform has all the desired properties (e.g., geometrical ratios and chemical composition) necessary to yield a high-quality fiber. The preform is subsequently drawn into a multi-kilometer-long hair-thin fiber. Four different preform manufacturing processes are currently in commercial use.

The most widely used process is the modified chemical vapor deposition (MCVD) process invented at the AT&T Bell Laboratories. Outside vapor deposition process (OVD) is used by Corning Glass Works and some of its joint ventures in Europe. Vapor axial deposition (VAD) process is the process used most widely in Japan. Philips, in Eindhoven, The Netherlands, uses a low-temperature plasma chemical-vapor deposition (PCVD) process.

In addition to the above four major processes, other processes are under development in different laboratories.

Plasma MCVD is under development at Bell Laboratories, hybrid OVD–VAD processes are being developed in Japan, and Sol-Gel processes are being developed in several laboratories. The first four processes are the established commercial processes and are producing fiber economically. The new processes are aimed at greatly increasing the manufacturing productivity of preforms and thereby reducing their cost.

All the above processes produce high-silica fibers using different dopants, such as germanium, phosphorous, and fluorine. These dopants modify the refractive index of silica, enabling the production of the proper core refractive-index profile. Purity of the reactants and the control of the refractive-index profile are crucial to the low loss and high bandwidth of the fiber.

1. MCVD Process

In the MCVD process (Fig. 8), a fused-silica tube of extremely high purity and dimensional uniformity is cleaned in an acid solution and degreased. The clean tube is mounted on a glass-working lathe. A mixture of reactants is passed from one end of the tube and exhaust gases are taken out at the other end while the tube is being rotated. A torch travels along the length of the tube in the direction of the reactant flow. The reactants include ultra-high-purity oxygen and a combination of one or more of the halides and oxyhalides ($SiCl_4$, $GeCl_4$, $POCl_3$, BCl_3, BBr_3, SiF_4, CCl_4, CCl_2F_2, Cl_2, SF_6, and $SOCl_2$).

The halides react with the oxygen in the temperature range of 1300–1600°C to form oxide particles, which are driven to the wall of the tube and subsequently consolidated into a glassy layer as the hottest part of the flame passes over. After the completion of one pass, the torch travels back and the next pass is begun. Depending on the type of fiber (i.e., multimode or single-mode), a barrier layer or a cladding consisting of many thin layers is first deposited on the inside surface of the tube. The compositions may include B_2O_3–P_2O_5–SiO_2 or F–P_2O_5–SiO_2 for barrier layers and SiO_2, F–SiO_2, F–P_2O_5–SiO_2, or F–GeO_2–SiO_2–P_2O_5 for cladding layers. After the required number of barrier or cladding layers has been deposited, the core is deposited. The core compositions depend on whether the fiber is single-mode, multimode, step-index, or graded-index multimode. In the case of graded-index multimode fibers, the dopant level changes with every layer to provide a refractive index profile that yields the maximum bandwidth.

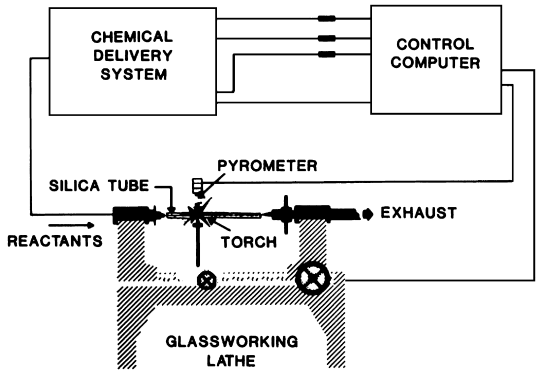


FIGURE 8 Schematic diagram of the MCVD process.

After the deposition is complete, the reactant flow is stopped except for a small flow of oxygen, and the temperature is raised by reducing the torch speed and increasing the flows of oxygen and hydrogen through the torch. Usually the exhaust end of the tube is closed first and a small positive pressure is maintained inside the deposited tube while the torch travels backward. The higher temperatures cause the glass viscosity to decrease, and the surface tension causes the tube to contract inward. The complete collapse of the tube into a solid preform is achieved in several passes. The speed of the collapse, the rotation of the tube, the temperature of collapse, and the positive pressure of oxygen inside the tube are all accurately controlled to predetermined values in order to produce a straight and bubble-free preform with minimum ovality. The complete preform is then taken off the lathe. After an inspection to assure that the preform is free of defects, the preform is ready to be drawn into a thin fiber.

The control of the refractive-index profile along the cross section of the deposited portion of the preform is achieved through a vapor delivery system. In this system, liquids are vaporized by passing a carrier gas (pure O_2) through the bubblers, made of fused silica. Accurate flows are achieved with precision flow controllers that maintain accurate carrier gas flows and extremely accurate temperatures within the bubblers. Microprocessors are used to automate the complete deposition process, including the torch travel and composition changes throughout the process. Impurities are reduced to very low levels by starting with pure chemicals, and there is further reducing of the impurities with in-house purification of these chemicals. Ultrapure oxygen and a completely sealed vapor-delivery system are used to avoid chemical contamination. Transition-metal ion impurities of well below 1 ppb and OH^- ion impurities of less than 1 ppm are typically maintained to produce high-quality fiber.

2. The PCVD Process

The PCVD process (Fig. 9) also uses a starting tube, and the deposition takes place inside the tube. Here, however, the tube is either stationary or oscillating and the

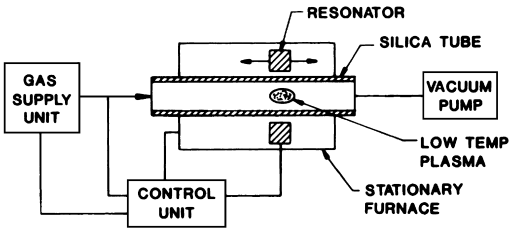


FIGURE 9 Schematic diagram of the PCVD process.

pressure is kept at 10–15 torr. Reactants are fed inside the tube, and the reaction is accomplished by a traveling microwave plasma inside the tube. The entire tube is maintained at approximately 1200°C. The plasma causes the heterogeneous depositions of glass on the tube wall, and the deposition efficiency is very high. After the required depositions of the cladding and core are complete, the tube is taken out and collapsed on a separate equipment. Extreme care is required to prevent impurities from getting into the tube during the transport and collapse procedure. The PCVD process has the advantages of high-efficiency, no-tube distortion because of the lower temperature and very accurate profile control because of the large number of layers deposited in a short time. However, going to higher rates of flow presents some difficulties because of a need to maintain the low pressure.

3. The PMCVD Process

The PMCVD is an enhancement of the MCVD process. Very high rates of deposition (up to 10 g/min compared to 2 g/min for MCVD) are achieved by using larger diameter tubes and an RF plasma for reaction (Fig. 10). Because of the very high temperature of the plasma, water cooling is essential. An oxyhydrogen torch follows the plasma and sinters the deposition. The high rates of deposition are achieved because of very high thermal gradients from the center of the tube to the wall and the resulting high thermophoretic driving force. The PMCVD process is still in the development stage and has not been commercialized.

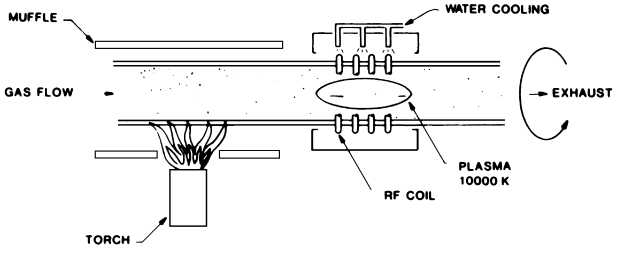


FIGURE 10 Schematic diagram of the PMCVD process.

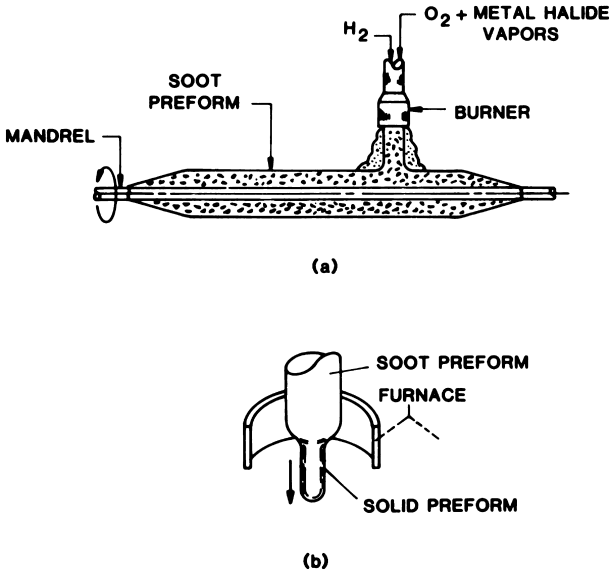


FIGURE 11 Schematic diagram of the outside vapor deposition. (a) Soot deposition. (b) Consolidation.

4. The OVD Process

The OVD process does not use a starting tube; instead, a stream of soot particles of desired composition is deposited on a bait rod (Fig. 11). The soot particles are produced by the reaction of reactants in a fuel gas–oxygen flame. A cylindrical porous soot preform is built layer by layer. After the deposition of the core and cladding is complete, the bait rod is removed. The porous preform is then sintered and dried in a furnace at 1400–1600°C to form a clear bubble-free preform under a controlled environment. The central hold left by the blank may or may not be closed, depending on the type of preform. The preform is now ready for inspection and drawing.

5. The VAD Process

The process is very similar to the OVD process. However, the soot deposition is done axially instead of radially. The soot is deposited at the end of a starting silica-glass rod (Fig. 12). A special torch using several annular holes is used to direct a stream of soot at the deposition surface. The reactant vapors, hydrogen gas, argon gas, and oxygen gas flow through different annular openings. Normally the core is deposited and the rotating speed is gradually withdrawn as the deposition proceeds at the end. The index profile is controlled by the composition of the gases flowing through the torch and the temperature distribution at the deposition surface. The porous preform is consolidated and dehydrated as it passes through a carbon-ring furnace in a controlled environment. $SOCl_2$ and Cl_2 are used to dehydrate the preform. Because of the axial deposition,

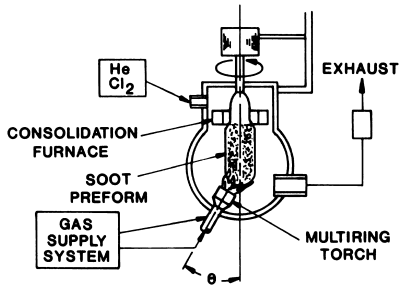


FIGURE 12 Schematic diagram of the vapor axial deposition.

this process is semicontinuous and is capable of producing very large preforms.

B. Fiber Drawing

After a preform has been inspected for various defects such as bubbles, ovality, and straightness, it is taken to a fiber-drawing station. A large-scale fiber-drawing process must repeatedly maintain the optical quality of the preform and produce a dimensionally uniform fiber with high strength.

1. Draw Process

During fiber drawing, the inspected preform is lowered into a hot zone at a certain feed rate, V_p, and the fiber is pulled from the softened neck-down region (Fig. 13) at a rate V_f. At steady state:

$$\pi D_p^2 V_p/4 = \pi D_f^2 V_f/4, \tag{1}$$

where D_p and D_f are the preform and fiber diameters, respectively. Therefore:

$$V_f = \left(D_p^2/D_f^2\right)V_p. \tag{2}$$

A draw machine, therefore, consists of a preform feed mechanism, a heat source, a pulling device, a coating device, and a control system to accurately maintain the fiber diameter and the furnace temperature.

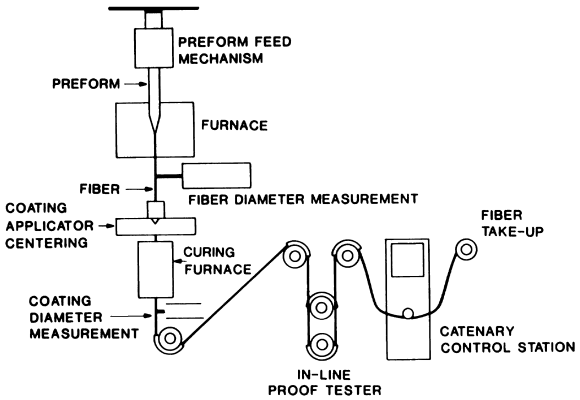


FIGURE 13 The fiber drawing process.

2. Heat Source

The heat source should provide sufficient energy to soften the glass for pulling the fiber without causing excessive tension and without creating turbulence in the neck-down region. A proper heat source will yield a fiber with uniform diameter and high strength. Oxyhydrogen torches, CO_2 lasers, resistance furnaces, and induction furnaces have been used to draw fibers. An oxyhydrogen torch, although a clean source of heat, suffers from turbulence due to flame. A CO_2 laser is too expensive a heat source to be considered for the large-scale manufacture of fibers. Graphite-resistance furnaces and zirconia-induction furnaces are the most widely used heat sources for fiber drawing. In the graphite-resistance furnace, a graphite-resistive element produces the required heat. Because graphite reacts with oxygen at high temperatures, an inert environment (e.g., argon) is maintained inside the furnace. The zirconia-induction furnace does not require an inert environment. It is extremely important that the furnace environment be clean in order to produce high-strength fibers. A zirconia-induction furnace, when properly designed and used, has produced very-high-strength long-length fibers (over 2.0 GPa) in lengths of several kilometers.

3. Mechanical Systems

An accurate preform feed mechanism and drive capstan form the basis of fiber speed control. The mechanism allows the preform to be fed at a constant speed into the hot zone, while maintaining the preform at the center of the furnace opening at the top. A centering device is used to position preforms that are not perfectly straight. The preform is usually held with a collet-type chuck mounted in a vertically movable carriage, which is driven by a lead screw. A precision stainless-steel drive capstan is mounted on the shaft of a high-performance dc servomotor. The fiber is taken up on a proper-diameter spool. The fiber is wound on the spool at close to zero tension with the help of a catenary control. In some cases fiber is proof-tested in-line before it is wound on a spool. The proof stress can be set at different levels depending on the application for which the fiber is being manufactured.

4. Fiber Coating System

The glass fiber coming out of the furnace has a highly polished pristine surface, and the theoretical strength of such a fiber is in the range of 15–20 GPa. Strengths in the range of 4.5–5.5 GPa are routinely measured on short fiber

lengths. To preserve this high strength, polymeric coatings are applied immediately after the drawing. The coating must be applied without damaging the fiber, it must solidify before reaching the capstan, and it should not cause microbending loss. To satisfy all these requirements, usually two layers of coatings are applied: a soft inner coating adjacent to the fiber to avoid microbending loss and a hard outer coating to resist abrasion. The coatings are a combination of ultraviolet- (UV) curable acrylates, UV-curable silicones, hot melts, heat-curable silicones, and nylons. When dual coatings are applied, the coated fiber diameter is typically 235–250 μm. The nylon-jacketed fiber typically used in Japan has an outside diameter of 900 μm. All coating materials are usually filtered to remove particles that may damage the fiber. Coatings are usually applied by passing the fiber through a coating cup and then curing the coating before the fiber is taken up by the capstan. The method of application, the coating material, the temperature, and the draw speed affect the proper application of a well-centered, bubble-free coating.

Fiber drawing facilities are usually located in a clean room where the air is maintained at class 10,000. The region of the preform and fiber from the coating cup to the top of the preform is maintained at class 100 or better. A class 100 environment means that there are no more than 100 particles of size greater than 0.5 μm in 1 ft^3 of air. A clean environment, proper centering of the preform in the furnace and the fiber in the coating cup, and proper alignment of the whole draw tower ensure a scratch-free fiber of a very high tensile strength. A control unit regulates the draw speed, preform feed speed, preform centering, fiber diameter, furnace temperature, and draw tension.

The coated fiber wound on a spool is next taken to the fiber measurement area to assure proper quality control.

5. Proof Testing of Fibers

Mechanical failure is one of the major concerns in the reliability of optical fibers. Fiber drawn in kilometer lengths must be strong enough to survive all of the short- and long-term stresses that it will encounter during the manufacture, installation, and long service life. Glass is an ideal elastic isotropic solid and does not contain dislocations. Hence, the strength is determined mainly by inclusions and surface flaws. Although extreme care is taken to avoid inhomogeneities and surface flaws during fiber manufacture, they cannot be completely eliminated. Since surface flaws can result from various causes, they are statistical in nature and it is impossible to predict the long-length strength of glass fibers. To guarantee a minimum fiber strength, proof testing has been adopted as a manufacturing step. Proof testing can be done in-line immediately after the drawing and coating or off-line before the fiber is stored.

In proof testing, the entire length of the fiber is subjected to a properly controlled proof stress. The proof stress is based on the stresses likely to be encountered by the fiber during manufacture, storage, installation, and service. The fibers that survive the proof test are stored for further packaging into cables.

Proof testing not only guarantees that the fiber will survive short-term stresses but also guarantees that the fiber will survive a lower residual stress that it may be subjected to during its long service life. It is well known that glass, when used in a humid environment, can fail under a long-term stress well below its instantaneous strength. This phenomenon is called static fatigue. Several models have been proposed to quantitatively describe the relationship between residual stress and the life of optical fibers. Use is made of the most conservative of these models, and the proof stress is determined by a consideration of the maximum possible residual stress in service and the required service life.

C. Fiber Packaging

In order to efficiently use one or more fibers, they need to be packaged so that they can be handled, transported, and installed without damage. Optical fibers can be used in a variety of applications, and hence the way they are packaged or cabled will also vary. There are numerous cable designs that are used by different cable manufacturers. All these designs, however, must meet certain criteria. A primary consideration in a cable design is to assure that the fibers in the cables maintain their optical properties (attenuation and dispersion) during their service life under different environmental conditions. The design, therefore, must minimize microbending effects. This usually means letting the fiber take a minimum energy position at all times in the cable structure. Proper selection of cabling materials so as to minimize differential thermal expansion or contraction during temperature extremes is important in minimizing microbending loss. The cable structure must be such that the fibers carry a load well below the proof-test level at all times, and especially while using conventional installation equipment. The cables must provide adequate protection to the fibers under all adverse environmental conditions during their entire service life, which may be as long as 40 years. Finally, the cable designs should be cost-effective and easily connectorized or spliced.

Five different types (Fig. 14) of basic cable designs are currently in use: (a) loose tube, (b) fluted, (c) ribbon, (d) stranded, and (e) Lightpack cable. The loose tube design was pioneered by Siemens in Germany. Up to 10 fibers are enclosed in a loose tube, which is filled with a soft filling compound. Since the fibers are relatively free to take the minimum energy configuration, the microbending losses

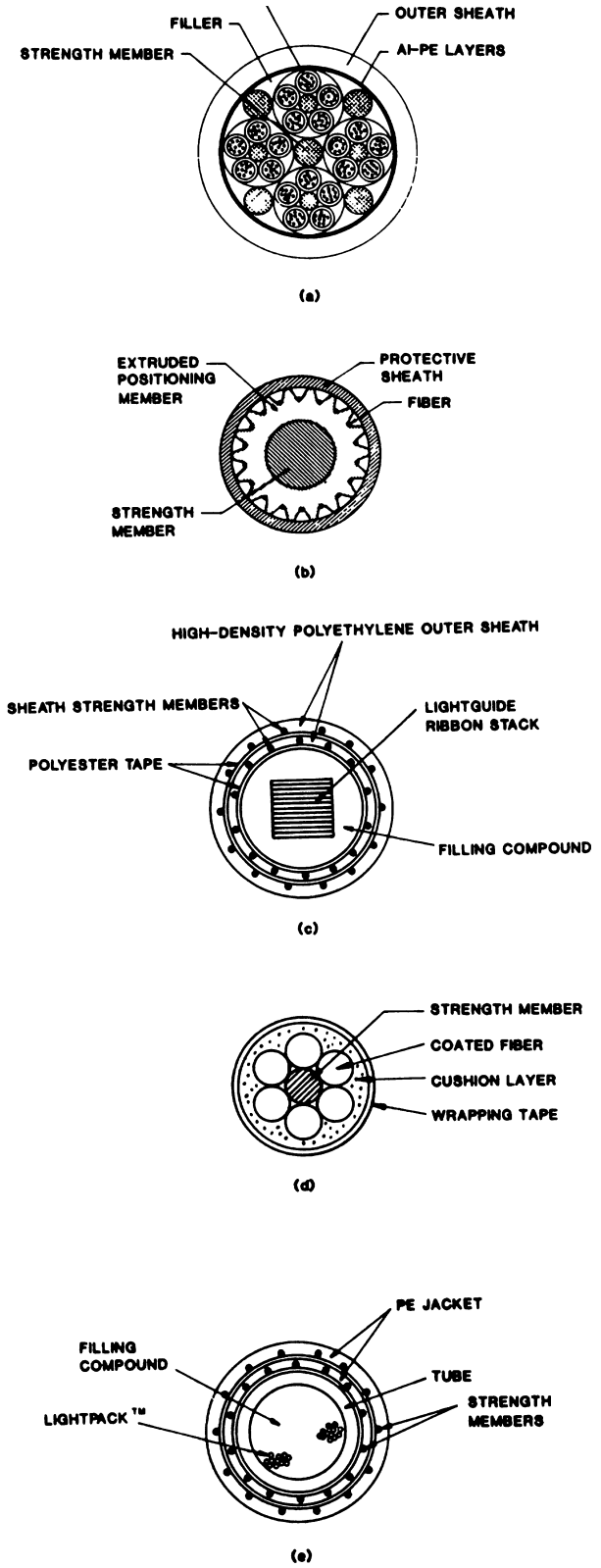


FIGURE 14 Fiber cable designs. (a) Loose tube design. (b) Slotted design. (c) Ribbon design. (d) Stranded unit. (e) Lightpack™ Cable design.

are avoided. Several of these buffered loose tube units are stranded around a central glass–resin support member. Aramid yarns are stranded on the cable core to provide strength members (for pulling through ducts), with a final polyethylene sheath on the outside. The stranding lay length and pitch radius are calculated to permit tensile strain on the cable up to the rated force and to permit cooling down to the rated low temperature without affecting the fiber attenuation.

In the fluted designs, fibers are laid in the grooves of plastic central members and are relatively free to move. The shape and size of the grooves vary with the design. The grooved core may also contain a central strength member. A sheath is formed over the grooved core, and this essentially forms a unit. Several units may then be stranded around a central strength member to form a cable core of desired size, over which different types of sheaths are formed. Fluted designs have been pioneered in France and Canada.

The ribbon design was invented at AT&T Bell Laboratories and consists of a linear array of 12 fibers sandwiched between two polyester tapes with pressure-sensitive adhesive on the fiber side. The spacing and the back tension on the fibers is accurately maintained. The ribbons are typically 2.5 mm in width. Up to 12 ribbons can be stacked to give a cable core consisting of 144 fibers. The core is twisted to some lay length and enclosed in a polyethylene tube. Several combinations of protective plastic and metallic layers along with metallic or nonmetallic strength members are then applied around the core to give the final cable its required mechanical and environmental characteristics needed for use in specified conditions. The ribbon design offers the most efficient and economic packaging of fibers for high-fiber-count cables. It also lends the cable to preconnectorization and makes it extremely convenient for installation and splicing.

The tight-bound stranded designs were pioneered in Japan and are used in the United States for several applications. In this design, several coated fibers are stranded around a central support member. The central support member may also serve as a strength member, and it may be metallic or nonmetallic. The stranded unit can have up to 18 fibers. The unit is contained within a plastic tube filled with a water-blocking compound. The final cable consists of several of these units stranded around a central member and protected on the outside with various sheath combinations.

The Lightpack cable design, pioneered by AT&T, is one of the simplest designs. Several fibers are held together with a binder to form a unit. One or more units are laid inside a large tube, which is filled with a water-blocking compound. This design has the advantage of the loose tube design in that the fibers are free of strain, but is

more compact. The tube-containing units can then be protected with various sheath options and strength members to provide the final cable.

The final step in cabling is the sheathing operation. After the fibers have been made into identifiable units, one or more of the units (as discussed earlier) form a core, which is then covered with a combination of sheathing layers. The number and combination of the sheathing layers depend on the intended use. Typically, a polyethylene sheath is extruded over the filled cable core. In a typical cross-ply design, metallic or nonmetallic strength members are applied over the first sheath layer, followed by another polyethylene sheath, over which another layer of strength members is applied. The direction of lay of the two layers of the strength members is opposite to each other. A final sheath is applied and the cable is ready for the final inspection, preconnectorization, and shipment. Metallic vapor barriers and lightning- and rodent-protection sheath options are also available. Further armoring is applied to cables made for submarine application.

In addition to the above cable designs, there are numerous other cable designs used for specific applications, such as fire-resistant cables, military tactical cables, cables for missile guidance systems, cables for field communications established by air-drop operations, air deployment cables, and cables for industrial controls. All these applications have unique requirements, such as ruggedness, low loss, and repeaterless spans, and the cable designs are accordingly selected. However, all these cable designs still rely on the basic unit designs discussed above.

V. SOURCES AND DETECTORS

In this section we review the optical sources and detectors that are used in fiber-optic communication systems.

A. Optical Sources

Semiconductor light-emitting diodes (LEDs) and injection-laser diodes (ILDs) are attractive as optical carrier sources because they are dimensionally compatible with optical fibers. They emit at wavelengths (0.8–0.9 μm and 1.3–1.6 μm) corresponding with regions of low optical-fiber loss, their outputs can be rapidly controlled by varying their bias current and therefore they are easy to modulate, and finally they offer solid-state reliability with lifetimes now exceeding 10^6 hr. Although LEDs and ILDs exhibit a number of similar characteristics, there are important differences between them that must be understood before one can select a source for a specific fiber-optic communication system.

One major difference between LEDs and ILDs is their spatial and temporal coherence. An ILD radiates a relatively narrow beam of light that has a narrow spectral width. In contrast, LED sources have much wider radiation patterns (beam width) and have moderately large spectral widths. These factors govern the amount of optical power that can be coupled into a fiber and the influence of chromatic dispersion on the bandwidth of the fiber medium. The second difference between ILDs and LEDs is their speed. The stimulated emission from lasers results in intrinsically faster optical rise and fall times in response to changes in drive current than can be realized with LEDs. The third difference between the devices is related to their linearity. LEDs generate light that is almost linearly proportional to the current passing through the device. Lasers, however, are threshold devices, and the lasing output is proportional to the drive current only above threshold. The threshold current of a laser, unfortunately, is not a constant but is a function of the device's temperature and age. Feedback control drive circuitry is therefore required to stabilize a laser's output power. Table IV illustrates typical ILDs and LED characteristics found in fiber-optic communication systems.

GaAlAs devices (both ILDs and LEDs) emitting in the wavelength region of 0.8–0.9 μm are commercially available and widely used in optical fiber systems. InGaAsP devices, with their emission wavelengths in the region 1.0–1.6 μm, are available for application near 1.3 and 1.6 μm, where fiber chromatic dispersion and transmission losses are minimal. The high-radiance Burrus-type (surface-emitting) LED is well suited for application in systems of low to medium bandwidth (<50 MHz). The power that can be coupled from an LED into a fiber is proportional to the number of modes the fiber can propagate, that is, to its core area times its numerical aperture squared. For simple butt coupling, where the emitting area of the LED is equal to or less than the core area of the fiber, presently available surface emitting GaAlAs and InGaAsP

TABLE IV Optical Source Characteristics

	ILDs	LEDs
Output power (mW)	1–10	1–10
Power launched into fiber (mW)	0.5–5	0.03–0.3
Spectral width (rms value) (nm)	2–4	15–60
Brightness (W/cm^2 sr)	$\sim 10^5$	10–10^3
Rise time, 10–90% (nsec)	<1	2–20
Frequency response (−3 dB) (MHz)	>500	<200
Voltage drop (V)	1.5–2	1.5–2.5
Forward current (mA)	10–300	50–300
Threshold current (mA)	5–250	Not applicable
Feedback stabilization required	Yes	No

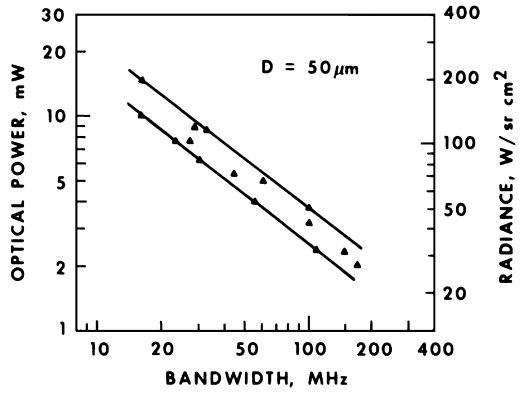


FIGURE 15 Output power versus 3-dB bandwidth for Burrus-type LED.

LEDs can launch about 50 μW into a graded-index fiber of numerical aperture (NA) 0.2 and core diameter 50 μm. The spectral width of an LED is a function of the operating wavelength, the active-layer doping concentration, and the junction current density. The rms spectral width of a typical 0.85-μm GaAlAs Burrus-type LED is about 16 nm, while that of a 1.3-μm InGaAsP LED is about 40 nm (spectral width is approximately proportional to λ^2). The modulation bandwidth of an LED depends on the device geometry, its current density, and the doping concentration of its active layer. Higher doping concentration yields higher bandwidth, but only at the expense of lower output power and wider spectral width. Figure 15 shows the trade-off between output power and modulation bandwidth for a group of Burrus-type GaAlSAs LEDs. Typically, a 50-μm-diameter LED that can butt-couple 50 μW into a 0.2-NA, 50-μm-core graded-index fiber can be current-modulated at rates up to about 50 MHz.

ILDs are well suited for application in medium- to high-bandwidth fiber-optic communication systems. Compared to LEDs, injection lasers offer the advantage of narrower spectral width (<3 nm), larger modulation bandwidth (>500 MHz), and greater launched power (1 mW). ILDs are the sources most compatible for use with single-mode fibers. However, ILDs are not as reliable as LEDs, are more expensive, and require feedback circuity to stabilize their output power against variations due to temperature and aging effects.

B. Optical Receivers–Photodetectors

The basic purpose of an optical receiver is to detect the received light incident on it and to convert it to an electrical signal containing the information impressed on the light at the transmitting end. The receiver is therefore an optical-to-electrical converter, or O/E transducer. An optical receiver consists of a photodetector and an associated amplifier along with necessary filtering and processing, as shown in Fig. 16. The function of the photodetector is to detect the incident light signal and convert it to an electrical current. The amplifier converts this current into a usable signal while introducing the minimum amount of additional noise to corrupt the signal. In designing an optical receiver, one tries to minimize the amount of optical power that must reach the receiver in order to achieve a given bit error rate (BER) in digital systems or a given signal-to-noise ratio (S/N) in an analog system. In this section we describe the characteristics of the photodetectors used in fiber-optic systems. Since the performance of an optical receiver depends not only on the photodetector but also on the components and design chosen for the subsequent amplifier, we also briefly describe configurations for this amplifier and their associated resulting receiver sensitivities.

In all the installed commercial fiber-optic communication systems in existence today, the photodetector used is either a semiconductor *p-i-n* or avalanche photodiode (APD). These devices differ in that the *p-i-n* basically converts one photon to one electron and has a conversion efficiency of less than unity. In an APD carrier, multiplication takes place that results in multiple electrons at the output per incident photon.

The reasons for choosing *p-i-n* or APD photodetector are usually based on cost and on required receiver sensitivity. The avalanching process in the APD has a sharp threshold, which is sensitive to ambient temperature and may require dynamic control of a relatively high bias voltage. The APD control and driver circuits are more expensive than those for the *p-i-n* detector, and the APD itself is more expensive than the *p-i-n* device. An APD with optimum gain, however, provides about 15 dB more receiver sensitivity than that achieved with a *p-i-n* diode.

An excellent spectral match exists between GaAlAs sources operating in the wavelength range of 0.8–0.9 μm and photodiodes made of silicon (spectral range 0.5–1.1 μm). The silicon *p-i-n* diode having no gain but with low dark current ($<10^{-9}$ A) and large bandwidth (1 GHz) is best suited for applications where receiver sensitivity is not critical. Silicon APDs are preferable in applications that demand high sensitivity, and those employed in presently installed telecommunication systems have current gains of about 100 and primary dark currents in the range of 10^{-10} to 10^{-11} A. Germanium photodiodes are used in the longer- wavelength region (1.3–1.6 μm), since the response of silicon decreases rapidly as λ increases beyond 1.0 μm. Germanium APDs with gain bandwidth products of approximately 60 GHz have been made, but their dark currents are high (10^{-8} to 10^{-7} A) and their excess noise factors are large. InGaAs and InGaAsP diodes are also used in the long-wavelength region. InGaAs

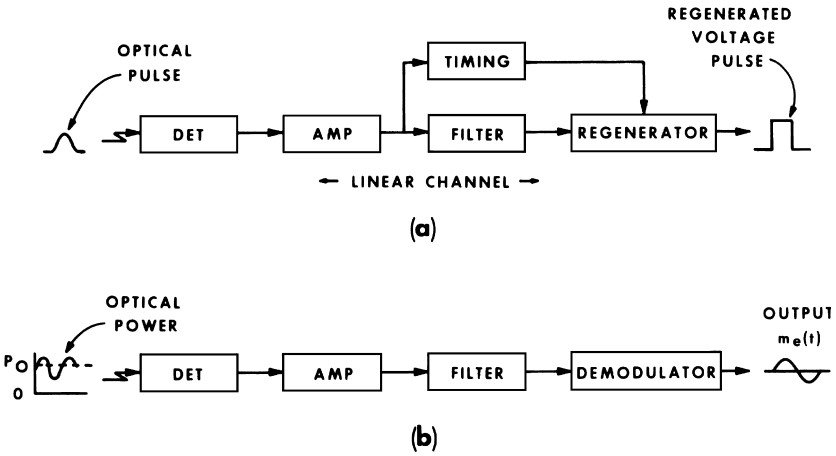


FIGURE 16 Schematic of digital receiver. (a) Digital. (b) Analog.

p-i-n diodes have been made with very low capacitance (<0.3 pF) and acceptably low dark current ($<5 \times 10^{-9}$ A). However, further work is required in the area of long-wavelength APDs to reduce their high excess noise factor. Table V summarizes the characteristics of photodetectors used in fiber-optic communication systems.

To calculate the system margin for a communication system, a knowledge of receiver sensitivity is needed. Receiver sensitivity is determined primarily by the characteristics of the photodetector and the low-noise front-end amplifier, which is optimized for use with the detector. Figure 17 shows two commonly used configurations. To achieve the best receiver sensitivity, the amplifier should have a high input impedance or provide feedback as in a transimpedance amplifier. The first stage can be either a GaAs field-effect transistor (FET) or a silicon bipolar transistor with a suitably adjusted emitter bias.

In the short-wavelength region (0.8–0.9 μm), silicon APDs can provide sufficiently high gain and low excess noise to overcome the input amplifier noise. In this wavelength region, design of the first amplifier stage is not very critical. A conventional silicon bipolar transistor having an emitter capacitance of 10 pF and a current gain of 150

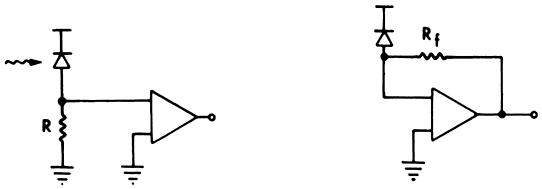


FIGURE 17 Schematic of amplifier designs.

has been used to build a digital receiver that requires only 10 nW average optical input power (−50 dBm) for a bit error rate (BER) of 10^{-9} at 100 Mbit/sec.

Since leakage currents (high noise factor) severely limit the use of avalanche gains as lownoise amplification process in the long-wavelength region (1.3–1.6 μm), *p-i-n* detectors are usually used. With a *p-i-n* diode, the microwave GaAs FET is well suited for use in the first amplifier stage because it has a low gate capacitance and high transconductance. It is typically used in a transimpedance configuration and offers wide bandwidth and good dynamic range. The best receivers have been built using InGaAs *p-i-n* detectors and GaAs FETs and require an average optical input power of 25 nW (−46 dBm) for 10^{-9} BER at 100 Mbit/sec.

TABLE V Typical Photodetector Characteristics

Characteristic	Silicon	Germanium	*p-i-n* Diodes		
			InP	Silicon	Germanium
Wavelength range (μm)	0.4–1.1	0.5–1.8	1.0–1.6	0.4–1.1	0.5–1.65
Wavelength of peak sensitivity (μm)	0.85	1.5	1.26	0.85	1.5
Quantum efficiency, 1%	80	50	70	80	70
Rise time (nsec)	0.01	0.3	0.1	0.5	0.25
Bias voltage (V)	15	6	10	170	40
Responsitivity (A/W)	0.5	0.7	0.4	0.7	0.6
Avalanche gain	1.0	1.0	1.0	80–150	80–150

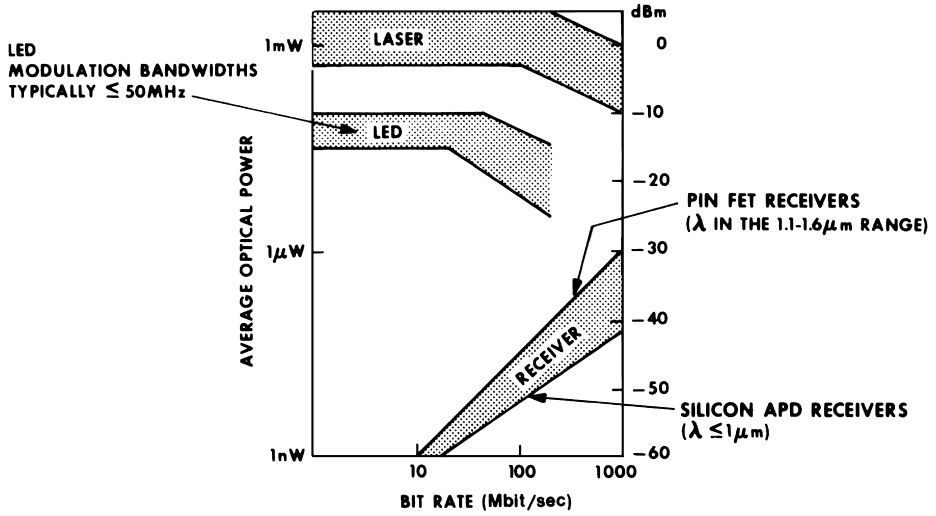


FIGURE 18 Transmission margin versus bit rate for digital transmission systems.

Curves for receiver sensitivity as a function of bit rate (bandwidth), along with curves of power available from LED and ILD sources, will allow us to calculate the net transmission loss tolerable between regenerators (system margin) as a function of bit rate (frequency).

Figure 18 shows, for digital systems, the average optical power required at the receiver (for BER of 10^{-9}) as well as the power available from optical sources as a function of bit rate. The lower boundary for receiver performance applies for receivers using silicon APDs at wavelengths of less than 1 μm. The upper receiver curve in Fig. 18 reflects the performance of *p-i-n* FET receivers sensitive in the wavelength range between 1.1 and 1.6 μm. Sources made of GaAlAs and InGaAsP have similar characteristics in terms of modulation speed and power delivered into a fiber. Light-emitting diodes are usually restricted to those applications where the modulation bandwidths are less than 50 MHz. The separation between the sources and receiver bands in Fig. 18 is an indication of the gross transmission margin of a system. Practical repeater spans are designed with about 10 dB subtracted from the maximum values given in Fig. 18. This will account for variation in the transmitter and receiver components due to temperature variations and aging. As a general rule of thumb, the required fiber bandwidth in a digital system is equal to or larger than the specified system bit rate. The noise penalty for using this rule is less than 1 dB. Along with this potential noise source, the 10-dB safety margin also allows for signal degradation from various noise sources in the transmitter and receiver. Once the net transmission loss that is tolerable between regenerators is obtained, the distance between regenerators (link length) can be determined from the loss characteristics for the fibers and interconnection devices used in the system.

VI. SOURCE COUPLING, SPLICES, AND CONNECTORS

In this section we investigate the coupling of energy from an optical source into a fiber and the effects of intrinsic and extrinsic splice-loss parameters on the transmission characteristics of an optical fiber link. In addition, we give examples of different types of optical fiber connectors and splices.

A. Source Coupling

The factors that effect the coupling efficiency of a source into a fiber can be broadly divided into two categories, as shown in Table VI. The first category, loss due to unintercepted illumination, can be caused by the source's emitting area being larger than the fiber's core area. Even if the source is smaller than the core, one can still have problems with unintercepted illumination if separation and misalignment of the source and fiber axes allow emitted light to miss the core and become lost. Coupling loss due to unintercepted illumination can be eliminated, however, if the source emitting area and the fiber-core area are properly matched and aligned. Figure 19 shows a fiber "pigtail" permanently mounted and properly aligned on a Burrus-type

TABLE VI Factors Influencing Source Coupling Efficiency into a Fiber

1. Unintercepted illumination loss
a. Area mismatch between source spot size and fiber core area
b. Misalignment of source and fiber axis
2. Numerical aperture loss, caused by that part of the source emission profile that radiates outside of the fiber's acceptance cone

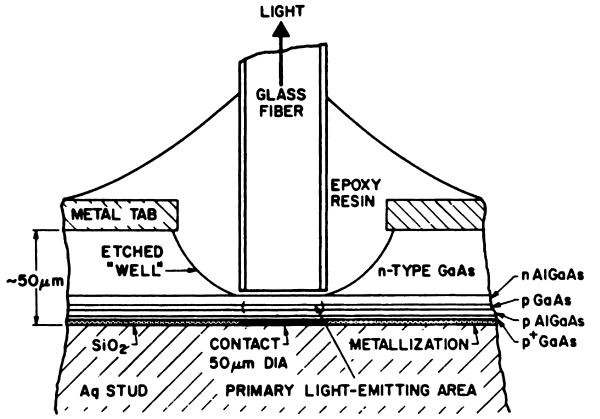


FIGURE 19 Burrus-type LED with a fiber pigtail. Small-area, high-radiance GaAs–Al_xGa_{1-x} As double-heretostructure surface emitter.

light-emitting diode. This configuration essentially eliminates loss due to unintercepted illumination. The second category of coupling loss that affects efficiency of source coupling into a fiber is due to mismatches between the source beam and fiber numerical apertures. This type of mismatch is shown schematically in Fig. 20. For fiber-optic communication systems, two types of light sources, light-emitting diodes (LEDs) and injection laser diodes (ILDs), are typically used. A Lambertian source whose radiant intensity varies with the cosine of the angle between a line perpendicular to it and another line drawn to an observation point is often used to model the radiation pattern of a surface-emitting LED. Some sources such as edge-emitting LEDs and ILDs exhibit radiation patterns that are narrower than a Lambertian source.

The radiation pattern obtained from an edge-emitting LED is elliptical in cross section with half-power beam divergence angles of approximately $\pm 60^\circ$ and $\pm 30^\circ$. The radiation pattern obtained from a double heterojunction laser diode is also elliptical in cross section but narrower in beam width than an LED. For example, the typical half-power beam divergence angles of an ILD are $\pm 25^\circ$ and $\pm 5^\circ$ perpendicular and parallel to the junction plane, respectively (see Fig. 21).

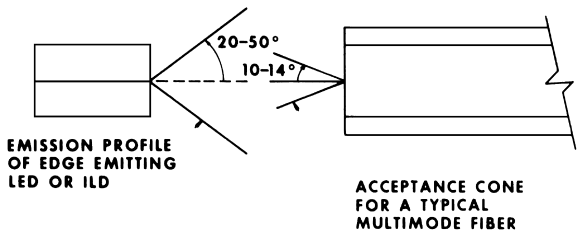


FIGURE 20 Schematic showing numerical-aperature mismatch of source and fiber.

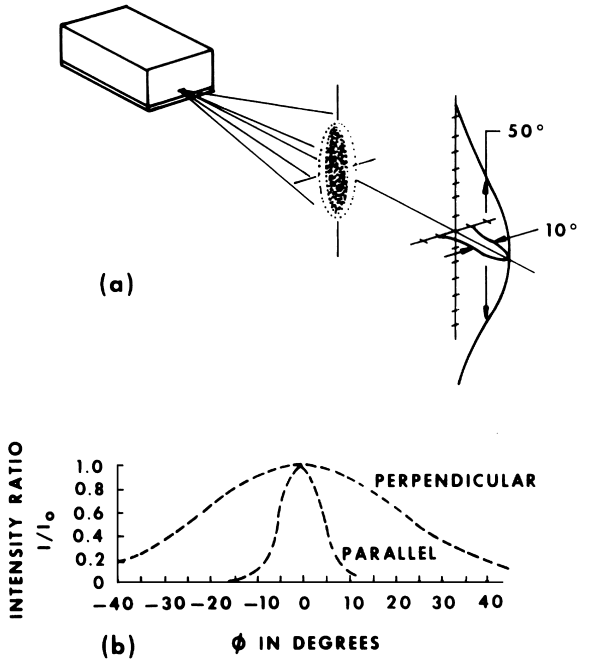


FIGURE 21 Radiation pattern of an injection laser diode. (a) Schematic representation of far-field radiation pattern. (b) Far-field intensity pattern measured in planes parallel and perpendicular to the junction.

If we consider the problem of coupling energy from an LED into a multimode graded-index fiber and assume that the LED is a Lambertian source in direct contact with the fiber core and covering its entire cross section, we can calculate the source coupling efficiency to be approximately equal to $\frac{1}{2}(\mathrm{NA})^2$. Sources such as edge-emitting LEDs and ILDs, which have more directional emitters, will have a higher coupling efficiency than a Lambertian source. ILDs with a properly aligned fiber pigtail will have a source coupling efficiency of 10–20%. For the case where the emitting area of the source is smaller than that of the fiber core, imaging optics can improve the source coupling efficiency. Figure 22 illustrates a number of different lens arrangements that have been used and their associated source coupling efficiencies. Efficiently coupling a light source into a single-mode fiber is a difficult task because of its small core diameter and small difference in the core cladding refractive index. Mechanical tolerances are more stringent for small-core fibers, and careful alignment is needed between the emitting area of the source and the fiber core. Table VII is an estimate for the power launched by surface- (SLED) and edge- (ELED) emitting LEDs and ILDs into both single-mode and multimode fibers.

B. Splices and Connectors

The practical implementation of optical fiber communication systems requires the use of interconnection devices

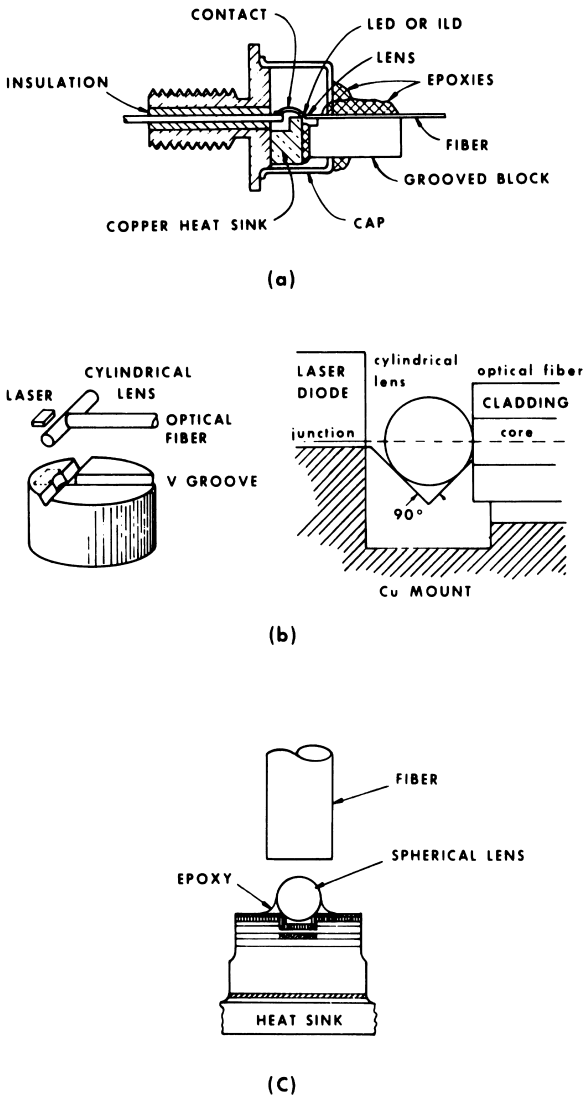


FIGURE 22 Lens system to increase source coupling efficiency. (a) LED or ILD with lens incorporated in the source package. (b) ILD with cylindrical lens coupling source energy into graded index fiber ($\eta_c \approx 40$–50%). (c) LED with spherical lens coupling source energy into fiber ($\eta_c \approx 8$–10%).

such as splices or connectors. A connector, by definition, is a demountable device used where it is necessary or convenient to easily disconnect and reconnect fibers. A splice, on the other hand, is employed to permanently join lengths of fiber together. The losses introduced by splices and connectors are an important factor to be considered in the design of a fiber-optic system, since they can be a significant part of the loss budget of a multikilometer communication link.

In this section we divide losses of splices and connectors into two categories, as shown in Table VIII. The first category of losses is related to the technique used to join fibers and is caused by extrinsic (to the fiber) parameters such as

TABLE VII Launched Power (in dB/min)

	Wavelength (μm)				
	0.85	1.3		1.55	
Diode	MMF[a]	MMF	SMF	MMF	SMF
SLED	−15	−17	−34	−18	−35
ELED	−9	−11	−22	−12	−23
ILD	+8	+5	+3	+4	+2

[a] MMF, multimode fiber; SMF, single-mode fiber.

transverse offset between the fiber cores, end separation, axial tilt, and fiber-end quality. The second category of losses is related to the properties of the fibers joined and is referred to as intrinsic (to the fibers) splice loss. Intrinsic parameters include variations in fiber diameter (both core and cladding), index profile (α and Δ mismatch), and ellipticity and concentricity of the fiber cores.

Figure 23 compares the relative influence on splice loss of the major extrinsic parameters of transverse offset, end separation, and axial tilt for graded-index multimode fibers. Splice loss is significantly more sensitive to transverse offset and axial tilt than it is to longitudinal offset. For example, a transverse offset of 0.14 core radii or an axial tilt of 1° (for a fiber with NA $= 0.20$) will produce a splice loss of 0.25 dB. A longitudinal offset of one core radius will produce a loss of only 0.14 dB. Fiber-end quality has a minimal effect on splice loss if proper fracturing or grinding and polishing end-preparation techniques are used in conjunction with an index-matching material. A matching material with a refractive index approximately the same as that of the core is used to reduce the Fresnel reflection loss caused by the glass–air interfaces between the coupled fibers of a joint.

Figure 24 compares, for single-mode fibers, the effects of transverse offset, end separation, and axial tilt on splice loss. Notice that very small (fractions of a micrometer) transverse offsets or axial tilts (fraction of a degree) can cause a significant amount of splice loss (>0.1 dB). The mismatch of intrinsic fiber parameters can

TABLE VIII Splice Loss Factors

Extrinsic
Transverse offset
Longitudinal offset
Axial tilt
Fiber end quality
Intrinsic
Fiber diameter variation
α Mismatch
Δ Mismatch
Ellipticity and concentricity of fiber core

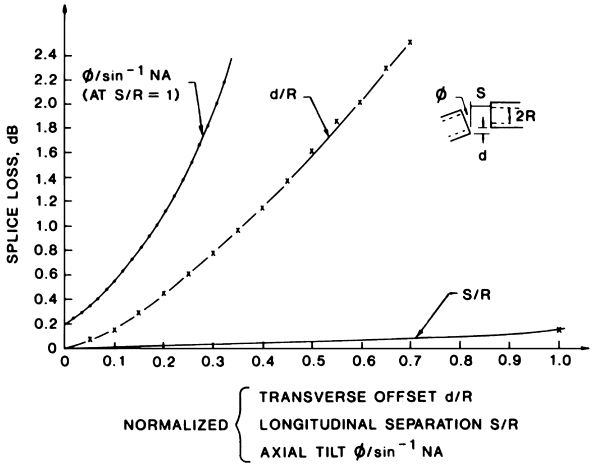


FIGURE 23 Splice loss due to extrinsic parameters.

also significantly affect the loss of a splice. For graded-index multimode fibers, splice loss is most sensitive to a mismatch of core diameter or delta. A normalized core radius or delta mismatch of 0.1 will produce a splice loss of approximately 0.2 dB. For single-mode fibers, the splice is most sensitive to a mismatch in mode field diameter. A mismatch in the mode diameter of 0.5 μm can cause a splice loss of 0.05 dB. To minimize the effect of intrinsic parameters on splice loss, tight manufacturing tolerances on the fibers used in a low-loss communication system must be maintained.

There is a variety of techniques that have been developed to interconnect fibers. Table IX describes some of the salient features of three of the techniques that are illustrated in this text.

Figure 25 is a good example of a passively aligned plug and alignment sleeve type of connector. This molded bi-

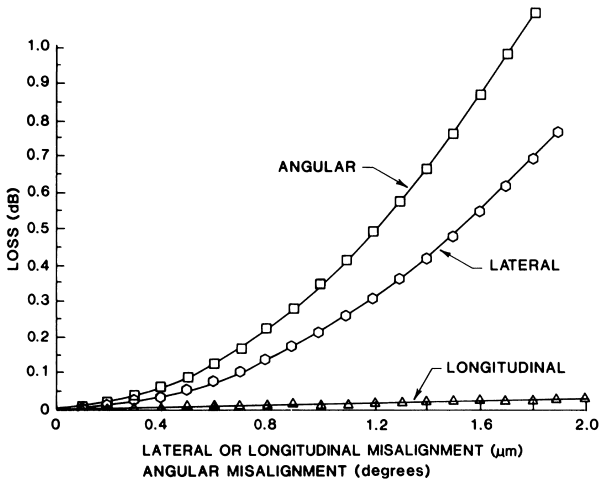


FIGURE 24 Extrinsic splice loss of single mode fibers.

conical plug connector is widely used, with both single-mode and multimode fibers, as part of a jumper cable for a variety of central-office applications. The heart of this connector is a biconical sleeve that accepts two plugs and aligns the axes of the fiber ends that are centrally located in these plugs. An inherent advantage of the conical alignment configuration is that virtually no abrasive wear occurs between the mating parts until the plug is fully seated within the biconical alignment sleeve. Losses of less than 1.0 dB are usually obtained with this type of connector. Another example of a plug and sleeve splice that illustrates the concept of active alignment of fibers is shown in Fig. 26. For single-mode splices, micrometer-type extrinsic and intrinsic tolerances must be maintained to achieve low-loss connections. To avoid putting very stringent constraints on both the fibers and the parts used to interconnect them, the technique of active alignment of the fiber cores while monitoring a signal related to transmission loss is used. Single-mode fibers mounted in precision glass cylindrical plugs are actively aligned in an eccentric sleeve to produce the very low loss (<0.1 dB) rotary splices shown in Fig. 26.

The silicon chip array splice shown in Fig. 27 is a very good example of the use of V grooves as the alignment mechanism for joining fibers. This splice is used in conjunction with ribbon-type structures to simultaneously splice groups of optical fibers (multiple-fiber splice). This reenterable splice consists of two array halves, two negative chips with metal backing, and two spring clips. An array half is formed by permanently affixing two positive preferentially etched silicon chips to the end of a ribbon. The fiber ends are then simultaneously prepared by grinding and polishing the end of the array. The splice halves are usually assembled on a cable prior to shipment from a factory, although they can be assembled in the field. A craftsperson assembles a splice by simply aligning two array halves with two negative chips. The splice is held together with spring clips, and a matching material is inserted to complete the connection. The reenterable array splice can be disconnected and reconnected in the field, although it was not designed for use where many connect–disconnect operations are required. The average loss of the silicon chip array splice is less than 0.2 dB for multimode fibers and less than 0.6 dB for single-mode fibers.

Fusion splicing, illustrated schematically in Fig. 28, is a widely used method for permanently joining individual fibers together. Fusion splicing, or welding as it is sometimes called, is accomplished by applying localized heating at the interface between two butted, realigned fiber ends, causing them to soften and fuse together. Commercially available fusion splicing machines produce splices with average losses of less than 0.3 dB (active

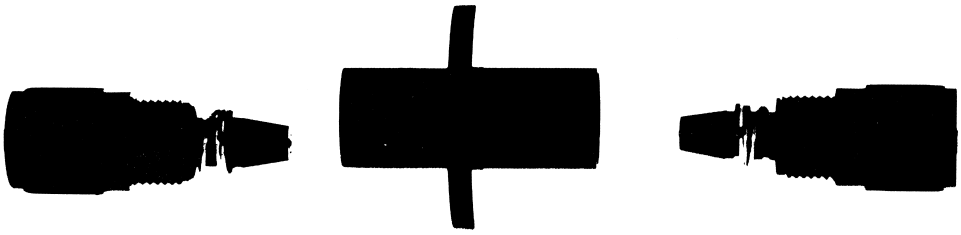
TABLE IX Characteristics of Interconnection Methods

Interconnection method	Alignment technique		Single-mode fiber	Multimode fiber	Connection	
	Active	Passive			Single	Multiple
Plug and alignment sleeve	X	X	X	X	X	
V-Groove alignment		X	X	X	X	X
Fusion	X	X	X	X	X	

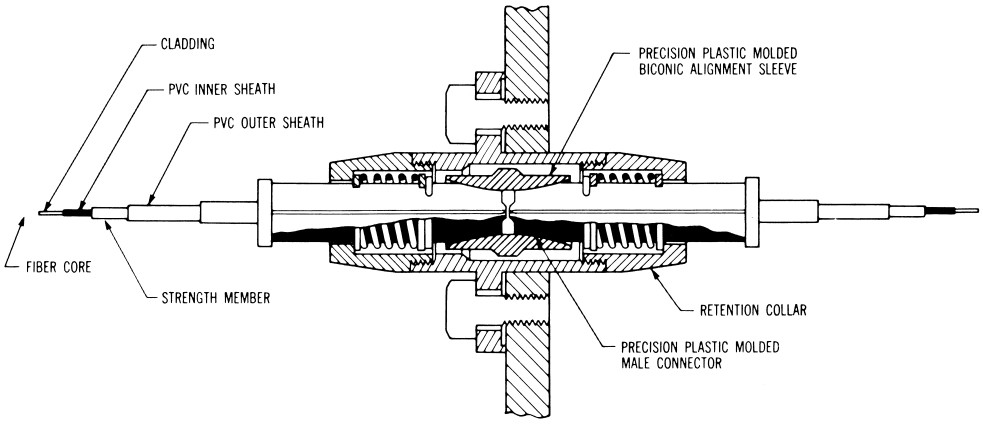
alignment) with single-mode fibers and less than 0.2 dB with multimode fibers.

VII. APPLICATIONS OF OPTICAL FIBER COMMUNICATIONS

The unique advantages of lightwave communications—the tremendous information-carrying capacity, freedom from electromagnetic interference, the light weight of fibers, and the relative chemical and high-temperature stability of fibers—suggest a communication revolution will occur in the near future. Already the telecommunications market for optical fibers has exploded in the United States, and unprecedented demand for fibers has been experienced. Within a short space of 5 years, from 1980 to 1985, the total U.S. circuit mile capacity installed with optical fibers far exceeded the total existing U.S. circuit mile capacity using copper cables in 1980. So far the field of fiber optics has been technology-driven, but the future will be application-driven. Lightwave application may be broadly classified into telecommunications applications, industrial applications, military application, medical applications, and sensor applications.



(a)



(b)

FIGURE 25 (a) Single-fiber biconical connector. (b) Cross section.

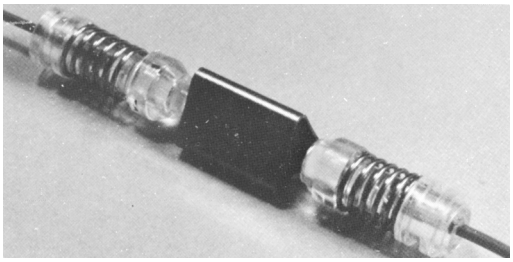


FIGURE 26 Rotary splice for single-mode fibers.

A. Telecommunications Applications

By far in terms of product volume and revenues the telecommunications application has been the major market. The need for greater circuit capacity coupled with the problem of congested duct space in the cities led to the initial applications of optical fiber communication for the interoffice trunk in the big cities (Fig. 29). However due to the tremendous progress made in reducing the transmission loss and increasing the bandwidth (information carrying capacity) of the optical fibers, they are now the economic choice for most telecommunication applications. Today the biggest market for fiber is the long-distance application. Japan has already completed its long-distance routes using single-mode fibers. In the United States, thousands of miles of lightguide cables are being installed by AT&T, MCI, GTE, and several other long-distance companies, as well as regional Bell and independent operating companies. The majority of these companies use single-mode fibers that operate at 1300 nm wavelength, with future upgradability to 1550 nm, when reliable sources and electronics are developed. Already commercial systems have been installed with bit rates as high as 2.4 Gbit/sec, which is equivalent to 30,720 two-way conversations over a pair of fibers. Another application for the fibers is in the feeder loop, that is, a cable that is used between a telephone central office and the local distribution interface. Since fibers have very high capacity and can transmit voice, data, and video, efforts are underway to install fibers into individual houses. AT&T has announced its intention to develop and offer a universal information service (UIS) to satisfy all the video, voice, and data communications needs of a subscriber. One of the most attractive features of the universal information system is the use of interactive systems. Systems are already in use that provide continuing interactive service for smoke and heat detectors to automatically alert fire departments of potential or active fire hazards, police alarms, and medical alert alarms to summon aid. In addition to the above, all the other information services can be obtained on demand. The potential for completely automating all the control needs of a household is there. Several trial projects have been implemented in Germany, France, Japan, the United Kingdom, Canada, and the United States. However, the information needs of individual households have not yet reached a level where fibers can be economically justified. However, definite potential exists for an information outlet in each household, which could be used for a multitude of services such as video phone, television, news, banking, and time-share computing. This is potentially one of the biggest future markets for fiber-optic applications.

B. Local Area Networks

Networks using optical fibers to transmit voice, video, and data within a building or within industrial and university campuses have been offered by several vendors. These systems are called local area networks (LANs). For economic reasons, many of today's LANs use a fiber with a high numerical aperture and operate with light-emitting diodes working at 1.30 μm. Fiber-optics LANs can improve the communications inside a high-use area, reduce the bulk of copper cables, and in many cases eliminate congestion in computer rooms by remoting peripheral equipment. LANs are likely to be installed in most office buildings, industrial facilities, university campuses, and military bases.

C. Industrial Applications

Industrial applications can be of different types, such as process control, manufacturing automation, and energy management and application. Applications for process control are in nuclear, petrochemical, chemical, and food industries. Applications for manufacturing automation are for numerical control of machines and large data systems. Applications for control exist in airways, highways, shipping, railways, gas and oil transportation, and distribution control. Although the information rates needed for such application are low compared to fiber capacity, fiber-optic control systems have several major advantages over the electronic wire systems. These include freedom from electromagnetic interference, improved signal quality, large potential for upgrading, and safety in a hazardous environment such as oil or gas. Also, remote control can be easily achieved. In the case of power failures, fiber-optic systems can be easily maintained with auxiliary power sources.

D. Computer Applications

Computer applications are the other part of industrial use of fiber optics. Fibers are ideally suited for internal links that require very high data rates, of the order of gigabits per second. Auxiliary equipments require lower data rates and hence can be handled both by fibers or copper wires; however, fibers offer the added advantage of longer distance

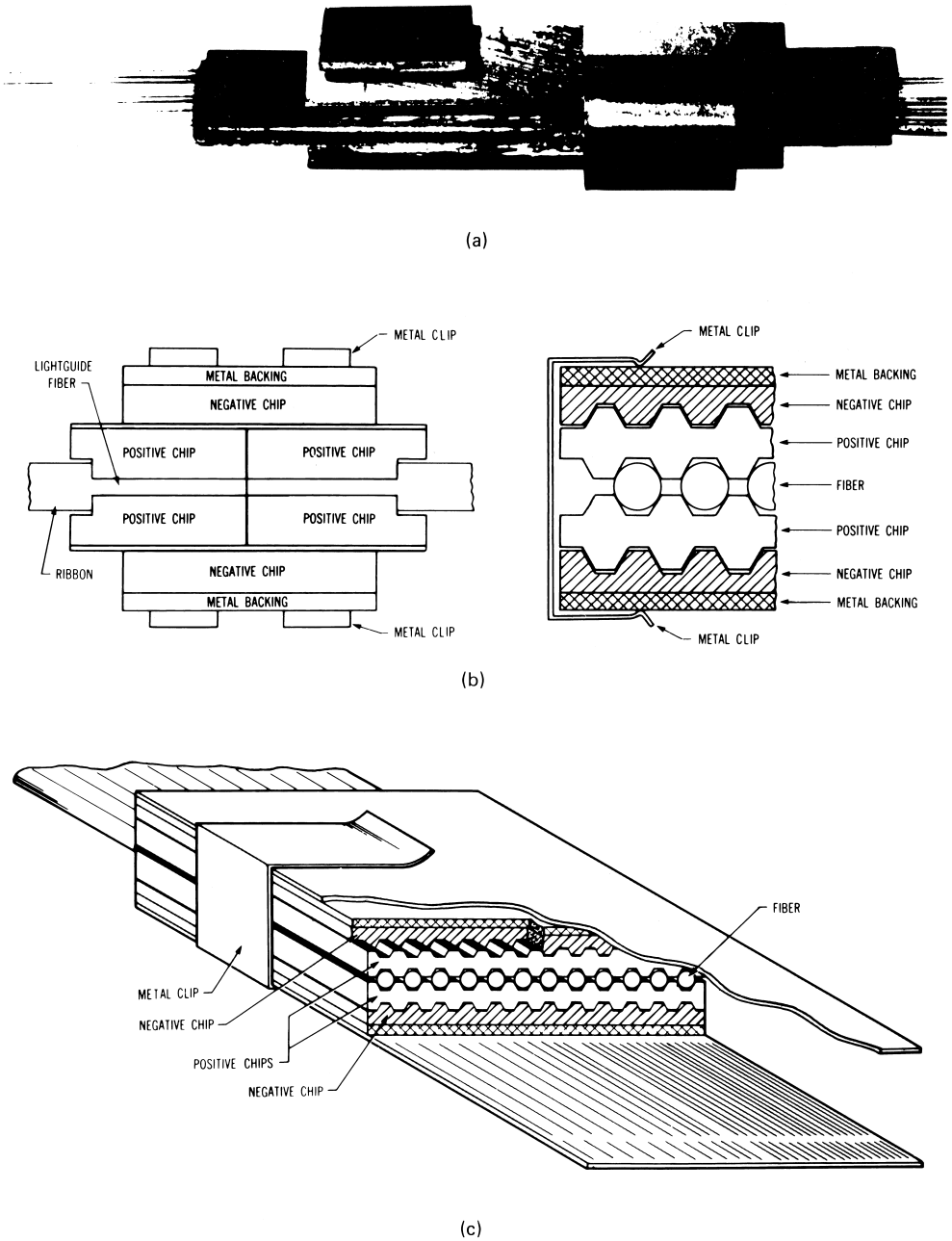


FIGURE 27 Multiple-fiber silicon chip array connector. (a) Twelve-fiber silicon chip array connector. (b) Side view and partial cross section of array connector. (c) Cross section of array connector.

network and error-free operation because fiber transmission is unaffected by the electromagnetic noise. The fibers will gradually be used in greater volume as inter- and intracomputer links.

E. Military Applications

Three attributes of fiber optics—security, freedom from electromagnetic interference, and light weight—make fibers very attractive for use in defense applications. Fibers can be used for data busing in aircrafts with tremendous reduction in weight, thereby increasing the range of these aircraft, along with increased information capacity and more secure and error-free operation. Such aircraft include surveillance aircraft, attack aircraft, space vehicles, and strategic air command bombers. Similar fiber applications exist for inter- and intraship communications, submarine mobile command centers, ship-to-satellite

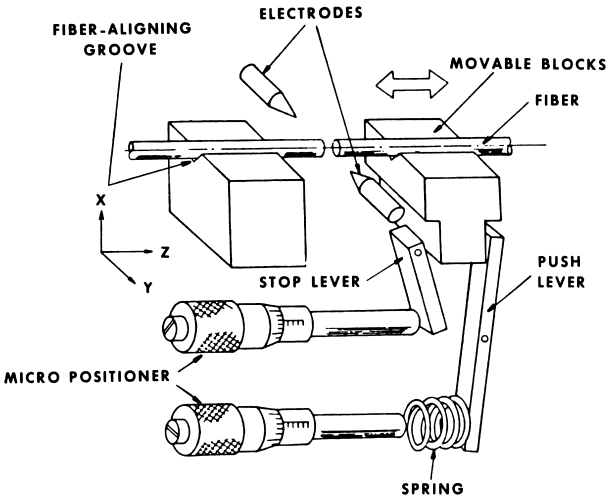


FIGURE 28 Schematic showing fusion splicing using an electric arc.

communication links, and all types of missile guidance systems. For military terrestrial applications, all communication within any base could be a local area network and could include missile–center links and ground-to-air and ground-to-ship communications. All the above types of applications exist today, making the military applications a large market for fiber-optics communications.

F. Medical Applications

In the medical field there are two distinctly different types of fiber-optics systems being offered. The first type takes the form of an LAN that will satisfy all the communications needs for a hospital, from keeping patient records to the energy-management and financial-management systems. The second category of medical fiber-optic systems uses fibers in diagnostic equipment and as sensors. Because of their small size and relatively benign impact upon body elements, much effort is underway to develop fiber-optic sensors for medical applications. The mechanical flexibility and small size allows the measurements of parameters inside the tissue, organs, and blood vessels. Also because of the freedom from electromagnetic interference, these instruments can be used in conjunction with other electronic monitoring systems. Three types of applications for biomedical diagnostics exists: (1) sensors for temperature, pressure, and flow of blood; (2) sensors for monitoring oxygen saturation, pH, and oxygen and carbon dioxide concentrations; and (3) sensors for immunoassay reactions. The sensors consist of fibers through which light is carried to the sample and returned back to analyzing equipment. At the end of the fiber is a suitable miniaturized transducer. The backscattered light is then analyzed to determine the velocity of moving blood cells using the Doppler shift. Spectral analysis can also be performed to measure concentration of blood constituents such as oxyhemoglobin. Concentration of gases in the bloodstream can be measured by coating the fiber end with an appropriate reagent for spectrophotometric or fluorimetric analysis. For monitoring immunoassay reactions, an antibody is attached to the surface of the fiber core from which cladding has been stripped. The reaction to the antigen or reaction product is measured at the fiber output end.

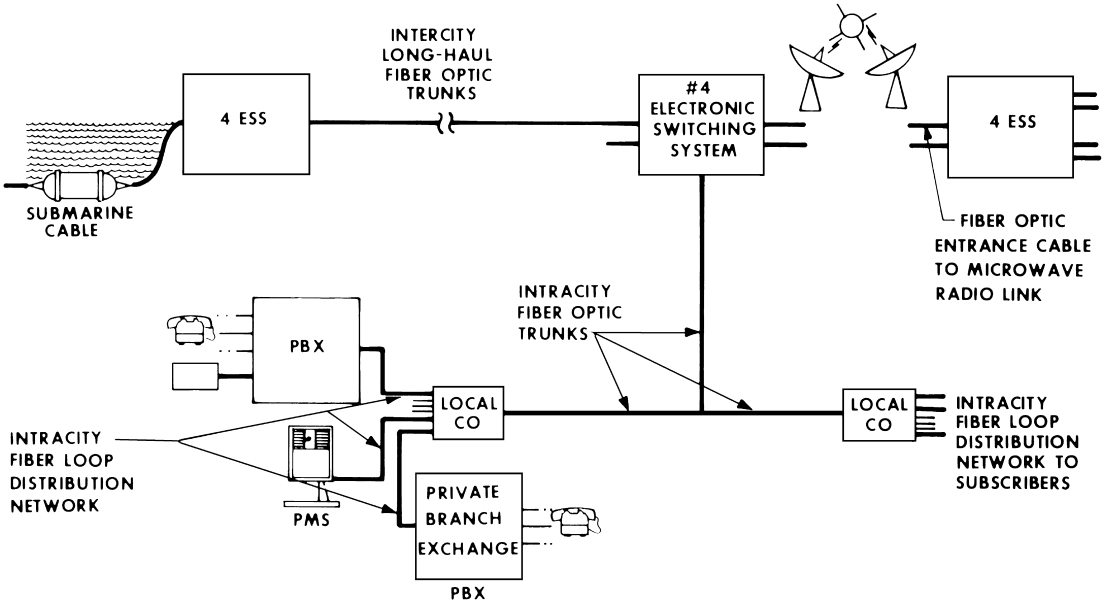


FIGURE 29 Optical fiber telecommunications application.

G. Sensor Applications

Use of fiber optics for sensors has been continuously increasing. As with other applications, the sensor applications also fall into three categories: military, medical, and industrial. The biomedical sensors were discussed in the previous section. Ultrasonic sensors use the phase shift caused in the light propagating through the core through radial stresses. The other sensors used are the pressure and temperature sensors, radiation sensors, depth sensors, acoustic sensors, sensors for color and turbidity, and sensors for underwater bubble detection. In general, the sensor system consists of an electronic control module, sensor head, and fiber-optic cable. The sensor head senses the pressure, temperature, velocity, and reaction and converts it into a change in the optical signal, which is then analyzed to measure the desired change by the electronic control unit. The field of fiber-optics sensors has expanded to the extent that separate conferences are held on fiber-optic sensors.

VIII. RECENT DEVELOPMENTS

Optical fiber communication has been one of the fastest developing technical and industrial fields in recent history. Already it is a multibillion-dollar industry worldwide, and it is one of the leading choices of careers for today's technical personnel. The amount of fiber deployed in the United States alone by various telecommunications companies was 7.2 million km at the end of 1989. Some of the recent developments in the field are summarized below.

A. Preform and Fiber Fabrication

Process and productivity improvements have resulted in optical fiber preforms yielding 200-km fibers routinely for the VAD and OVD processes and 25 km for the MCVD process. This has been achieved through increased deposition rates in these processes. Along with the productivity increases, there have been process improvements which have resulted in improved fiber performance. Today, fibers with attenuations of 0.34 dB/km at 1.31 μm and 0.19 dB/km at 1.55 μm are routinely manufactured. There has also been an increasing production and use of dispersion-shifted fibers, in which the zero dispersion wavelength is shifted to the 1.55-μm window. Hermetically sealed fibers with a-few-atoms-thick carbon coatings have been developed by Corning, AT&T, and Sumitomo. These fibers show static fatigue constants of more than 100 and are ideally suited for underwater applications.

The search for new types of fibers continues, and one of the most promising fibers is the erbium-doped fiber. These fibers can amplify signals and can perform as optical amplifiers.

B. System Performance

The availability of extremely low-loss fibers, dispersion-shifted fibers, and new lasers has led to the design of ultra high performance systems in terms of high bit rates and long repeaterless spans. NTT of Japan is installing a high-capacity backbone system using dispersion-shifted fibers.

AT&T, using erbium-doped fiber amplifiers and frequency shift key (FSK) transmission has delivered 6.8 Gbit/sec over a 70-km span in a typical field environment. Transmission of a 2.488 Gbit/sec signal over 2200 km with 25 in-line amplifiers has been demonstrated by NTT in Japan. AT&T Bell Laboratories has reported transmission of solitons over 6000 km of fiber without external amplification. A soliton is a nondispersive pulse that makes use of fiber nonlinearity to cancel out the effects of fiber's chromatic dispersion.

C. Applications

Three new areas of optical communications are expected to experience major growth in the future. These are Broadband Integrated Services Digital Networks (BISDN), fiber-to-the-home, and cable TV. A few ISDNs have been installed by the telephone companies for some of their large customers. These networks offer many different types of services, ranging from voice to data and video over one integrated network. Current networks, however, are not broadband because of a lack of demand. However, it is expected that the demand for services will greatly increase. Fiber-to-the-home is another major area of expected growth, and current projections are that this segment will grow steadily in the next decade. With respect to cable TV, optical fibers are already well established in the trunk, i.e., between central offices. The feeder plant also is utilizing optical fibers for connecting digital loop carriers to the central office. The last portion of the cable TV network, the distribution plant, is expected to fiberized in the near future.

SEE ALSO THE FOLLOWING ARTICLES

FIBER-OPTIC-BASED SMART STRUCTURES • FIBER-OPTIC CHEMICAL SENSORS • GLASS • LASERS, OPTICAL FIBER • OPTICAL FIBERS, FABRICATION AND APPLICATIONS • OPTICAL FIBER TECHNIQUES FOR MEDICAL APPLICATIONS

BIBLIOGRAPHY

Barnoski, M. K., ed. (1976). "Fundamentals of Optical Fiber Communications," Academic Press, New York.

Bendow, B., and Shashanka, S. M., ed. (1979). "Fiber Optics: Advances in Research and Development," Plenum Press, New York.

Cherin, A. H. (1983). "Introduction to Optical Fibers," McGraw–Hill, New York.

Li, T., ed. (1985). "Optical Fiber Communications," Academic Press, New York.

Midwinter, J. E. (1979). "Optical Fibers for Transmission," Wiley, New York.

Miller, C. M. (1986). "Optical Fibers Splices and Connectors," Marcel Dekker, New York.

Miller, S. E., and Chynoweth, A. G., ed. (1979). "Optical Fiber Telecommunications," Academic Press, New York.

Murata, H. (1988). "Handbook of Optical Fibers and Cables," Marcel Dekker, New York.

Suematsu, Y., and Ken-ichi, I. (1982). "Introduction to Optical Fiber Communication," Wiley, New York.



Optical Information Processing

Mir Mojtaba Mirsalehi
University of Alabama, Huntsville

I. History of Optical Information Processing
II. Background
III. Coherent Optical Processing
IV. Incoherent Optical Processing
V. Hybrid Processors
VI. Optical Digital Computing
VII. Applications

GLOSSARY

Bistability Property of having two stable output states for a given input. In optical bistable devices, the input and the output are beams of light, and the two states are normally the low and high transmittances of a medium.

Diffraction Deviation of light rays from rectilinear paths that cannot be interpreted as reflection or refraction.

F number Ratio between the diameter of a lens and its focal length.

Holography Technique for recording and reconstructing information about both amplitude and phase of light scattered from an object.

Index of refraction Ratio of the speed of light in a vacuum to the speed of light in the material.

Monochromatic light Radiation at a single frequency.

Photorefractive material Material whose index of refraction can be altered by exposing it to light.

Pupil function Function that describes the finite extent of a lens or an object.

Space–bandwidth product Product of the diameter of the input aperture of an opitcal system and the highest spatial frequency over which that system can operate.

Spatial filtering Manipulation of the spectrum of an image by a filter plate.

Spatial light modulator Device that manipulates the amplitude or/and the phase of a beam of light at different points of space.

OPTICAL INFORMATION PROCESSING involves processing a two-dimensional array of data using light. The main advantage of optical processors is their ability to process data in parallel (i.e., an array of data points is processed simultaneously). However, most of the optical systems are analog in nature and thus suffer from lack of accuracy and flexibility. Recently, hybrid (optical analog/digital electronic) systems have been developed to overcome this shortcoming. Significant research is currently underway to develop optical digital computing systems.

I. HISTORY OF OPTICAL INFORMATION PROCESSING

One of the earliest contributions in the field of optical information processing is Abbe's theory of microscope imagery, which was introduced in 1873. He showed that the image of an object might lose its detail or obtain false detail if not all parts of the diffracted light are transmitted through the optical system. Porter demonstrated this point by a remarkable experiment in 1906 (Fig. 1), in which lens L_1 collimates the light and L_2 provides an image of the object at P_i. Porter used a piece of wire gauze as the object. The Fourier transform of the object is formed at the back focal plane (P_f). If all parts of the spectrum are allowed to pass, a faithful image of the gauze is obtained. Porter showed that if a horizontal slit that allows only one row of the spectrum to pass is put at P_f, the vertical structure of the image is lost. Similarly, the horizontal structure is lost if a vertical slit is used (Fig. 2). This experiment is the first reported demonstration of spatial filtering.

In 1935, a significant advancement in the field of microscopy was realized with Zernike's phase-contrast microscope. This Nobel prize-winning invention made the observation of transparent objects possible. In 1946, the use of the Fourier integral in analyzing optical systems was introduced by Duffieux. In the early 1950s, Maréchal and his coworkers at Paris University successfully demonstrated a variety of coherent spatial filtering techniques to improve the quality of photographs. Also in that decade, a fruitful exchange of knowledge was begun between scientists working in communications and those in optics. The resemblance of many problems in optics to concepts in communications such as optimum filtering, detection, and estimation was shown by Elias and O'Neill. In the 1960s, coherent optical processing was successfully applied in the field of radar signal processing. Much of this progress was made by Cutrona and his coworkers at the University of Michigan Radar Laboratory. Two other important events in the field of optical information processing that occurred during the 1960s are the inventions of holographic spatial filtering by Vander Lugt in 1963 and computer-generated holograms for spatial filtering by Lohmann and Brown in 1966. In the 1970s, the field of optical processing was broadened by progress in coherent and incoherent processing and the development of hybrid systems. Also, stimulated by advances in technology and the rediscovery of residue arithmetic, the idea of building an optical digital computer was reborn in the mid-1970s. Significant research and development in all of the areas mentioned above is currently being pursued. The area of optical digital computing seems especially promising with new architectures for the future computers.

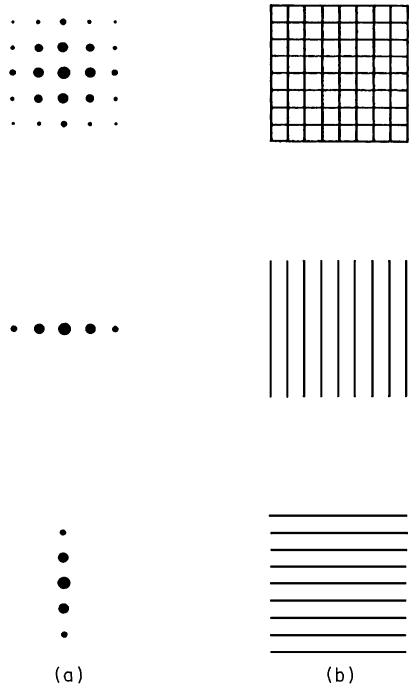


FIGURE 2 Results from Porter's experiment: (a) the spectra at the focal plane and (b) the corresponding images at the image plane.

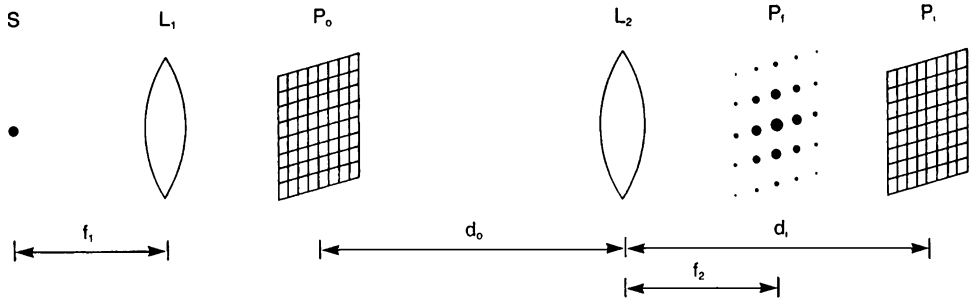


FIGURE 1 Porter's experiment, where d_i is the image distance; d_o, object distance; f_1, focal length of L_1; f_2, focal length of L_2; L_1, collimating lens; L_2, imaging lens; P_f, focal plane; P_i, image plane; P_o, object plane; and S, light source.

II. BACKGROUND

The analysis of optical systems requires some background knowledge. The fundamental concepts needed are reviewed in this section. This material is divided into three subsections: mathematics, system theory, and optics. Because of space constraints, the concepts are described briefly. References are cited in the bibliography for further information.

A. Relevant Mathematics

1. Special Functions

One of the functions used in analyzing optical systems is the Dirac delta (impulse) function $\delta(x)$. This function has the following properties:

$$\delta(x) = 0, \qquad x \neq 0 \tag{1a}$$

$$\int_{x_1}^{x_2} \delta(x)\, dx = 1, \qquad x_1 < 0 < x_2. \tag{1b}$$

A point source of light can be modeled as a two-dimensional delta function, since it is radiation of a finite energy through a small aperture. A periodic array of impulse functions is known as a comb function.

Among other functions used in optics are the rectangle, triangle, Gaussian, and sinc functions. These are, respectively, defined as follows:

$$\text{rect}(x) = \begin{cases} 1, & |x| \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$\Lambda(x) = \begin{cases} 1 - |x|, & |x| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\text{Gaus}(x) = \exp(-\pi x^2) \tag{4}$$

$$\text{sinc}(x) = \sin(\pi x)/\pi x. \tag{5}$$

These functions can be extended to two dimensions. Another useful two-dimensional function is the circle function, which is defined as follows:

$$\text{circ}(r) = \begin{cases} 1, & r = (x^2 + y^2)^{1/2} \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The above functions are shown in Fig. 3.

2. Convolution and Correlation Operations

The convolution of two real or complex functions $\mathbf{f}(x)$ and $\mathbf{g}(x)$ may be represented by $\mathbf{f}(x) * \mathbf{g}(x)$ and is defined as follows:

$$\mathbf{f}(x) * \mathbf{g}(x) = \int_{-\infty}^{+\infty} \mathbf{f}(\alpha)\, \mathbf{g}(x - \alpha)\, d\alpha. \tag{7}$$

Convolution is a commutative operation [i.e., $\mathbf{f}(x) * \mathbf{g}(x) = \mathbf{g}(x) * \mathbf{f}(x)$]. The result of convolving two functions can be considered as the area under the curve that is obtained by multiplying one function with a reversed (in α) and shifted (by x) version of the other function. Convolution is a useful operation in analyzing linear shift-invariant (LSI) systems as is shown later. Most optical systems can be considered two-dimensional LSI systems. The two-dimensional version of Eq. (7) is as follows:

$$\mathbf{f}(x, y) * \mathbf{g}(x, y) = \int_{-\infty}^{+\infty}\int \mathbf{f}(\alpha, \beta) \times \mathbf{g}(x - \alpha, y - \beta)\, d\alpha\, d\beta. \tag{8}$$

The complex cross-correlation operation, denoted by $\bigstar$, is defined as follows:

$$\mathbf{f}(x) \bigstar \mathbf{g}(x) = \int_{-\infty}^{+\infty} \mathbf{f}(\alpha)\, \mathbf{g}^*(\alpha - x)\, d\alpha, \tag{9}$$

where $\mathbf{g}^*$ represents the complex conjugate of $\mathbf{g}$. Although the cross-correlation and convolution operations look similar [in fact, $\mathbf{f}(x) \bigstar \mathbf{g}(x) = \mathbf{f}(x) * \mathbf{g}^*(-x)$], they differ in nature. Unlike convolution, cross-correlation is not commutative [i.e., $\mathbf{f}(x) \bigstar \mathbf{g}(x) \neq \mathbf{g}(x) \bigstar \mathbf{f}(x)$]. A special case of cross-correlation is called autocorrelation, when the two functions are the same [i.e., $\mathbf{f}(x) = \mathbf{g}(x)$]. Autocorrelation and cross-correlation are useful operations in pattern recognition.

3. Integral Transforms

An integral transform is a linear operation that converts a function, $\mathbf{f}(x)$, to another function, $\mathbf{F}(u)$, via the following integral:

$$\mathbf{F}(u) = \int_a^b \mathbf{f}(x)\mathbf{K}(x, u)\, dx. \tag{10}$$

The function $\mathbf{K}(x, u)$, known as the kernel of the transform, and the limits of the integral are specified for a particular transform. Integral transforms are used to map one domain into another in which the problem is simpler to analyze. For example, the analysis of linear time-invariant systems usually becomes easier if the time domain representation is changed to the frequency domain representation using the Fourier transformation.

A number of integral transforms, each suitable for different problem areas, exist. The most common transform in optics is Fourier transform. The Fourier transform of a one-dimensional function, $\mathbf{f}(x)$, is defined as follows:

$$\mathbf{F}(u) = \int_{-\infty}^{+\infty} \mathbf{f}(x) \exp(-j2\pi ux)\, dx, \tag{11}$$

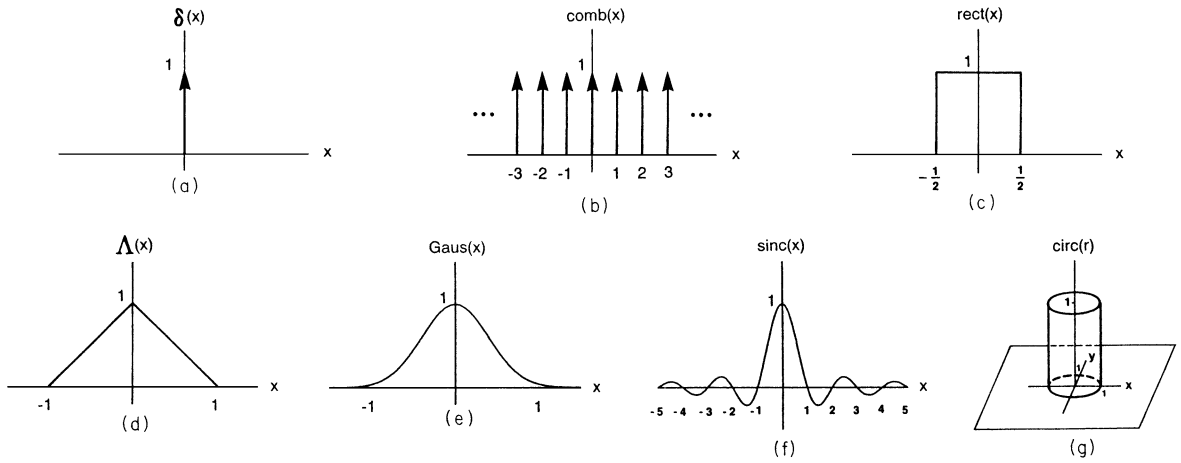


FIGURE 3 Some functions frequently used in optics: (a) impulse function, (b) comb function, (c) rectangle function, (d) triangle function, (e) Gaussian function, (f) sinc function, and (g) circle function.

where $j = (-1)^{1/2}$. The corresponding inverse transform is as follows:

$$\mathbf{f}(x) = \int_{-\infty}^{+\infty} \mathbf{F}(u) \exp(j2\pi ux)\, du. \tag{12}$$

One of the properties of the Fourier transform is known as the convolution theorem. According to this property, the Fourier transform of the convolution of two functions is the product of their Fourier transforms as follows:

$$\mathcal{F}\{\mathbf{f}(x) * \mathbf{g}(x)\} = \mathcal{F}\{\mathbf{f}(x)\}\,\mathcal{F}\{\mathbf{g}(x)\}, \tag{13}$$

where $\mathcal{F}$ represents the Fourier transform operation. This is a useful tool in analyzing LSI systems.

Most optical systems are two-dimensional, hence, they require two-dimensional transformations. The two-dimensional versions of Eqs. (11) and (12) can be written as follows:

$$\mathbf{F}(u, v) = \int_{-\infty}^{+\infty}\int \mathbf{f}(x, y) \exp[-j2\pi(ux + vy)]\, dx\, dy \tag{14}$$

and

$$\mathbf{f}(x, y = \int_{-\infty}^{+\infty}\int \mathbf{F}(u, v) \exp[j2\pi(ux + vy)]\, du\, dv. \tag{15}$$

It is common to replace u and v in these equations with f_X and f_Y which denote the spatial frequencies along the x and y axes.

Another integral transform of interest in optics is the Mellin transform, which is defined as follows:

$$\mathbf{F}(\mathbf{s}) = \int_0^{\infty} \mathbf{f}(x) x^{\mathbf{s}-1}\, dx, \tag{16}$$

where the parameter $\mathbf{s}$ is a complex variable. If for some $k > 0$, the inequality $\int_0^{\infty} x^{k-1}|\mathbf{f}(x)|\, dx < \infty$ is satisfied, the inverse Mellin transform can be obtained for any $a > k$ using the following equation:

$$\mathbf{f}(x) = \frac{1}{2\pi j} \int_{a-j\infty}^{a+j\infty} \mathbf{F}(\mathbf{s}) x^{-\mathbf{s}}\, d\mathbf{s}. \tag{17}$$

An interesting property of the Mellin transform is that it is scale-invariant. In other words, if the input function is stretched, the shape of the transform remains constant. This property can be stated as $\mathcal{M}\{\mathbf{f}(bx)\} = b^{-\mathbf{s}}\,\mathcal{M}\{\mathbf{f}(x)\}$, where $\mathcal{M}$ represents the Mellin transform and b is a constant.

4. Boolean Algebra

Boolean algebra was introduced by George Boole in 1847 as a tool for the mathematical analysis of logic. It was not used for practical purposes, however, until the late 1930s, when A. Nakashima and, independently, C. E. Shannon used it for analyzing relay contact networks. After World War II, the switching theory was extended to include sequential systems (i.e., systems whose outputs depend not only on the inputs but also on the previous states of the system). Based on these advances in theory, digital electronic computers and other digital systems were developed. Today, digital electronic systems are widely used, and the corresponding theories are studied in mathematics, computer science, and electrical engineering. The purpose of this subsection is to define only the basic terms used in the analysis of digital systems, so that a reader without a background in this field can understand the material presented in the optical digital computing section.

A digital system is defined as a system whose inputs and outputs are discrete quantities. These quantities

TABLE I A Truth Table

X	Y	Z	F
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

are normally represented by binary variables. A binary variable has two possible values: 0 and 1 (also known as false and true). A special case of Boolean algebra that deals with binary variables is known as switching algebra. A function that accepts binary inputs and provides binary outputs is called a binary function. A binary function can be defined by a corresponding truth table that provides the output for all possible combinations of values of the input variables. Table I shows a truth table that corresponds to a function with three input variables (X, Y, and Z) and one output variable (F).

Another method of defining a binary function is to express each output variable in terms of the input variables using some of the basic logic operations. The basic logic operations include NOT, AND, OR, NAND, NOR, and EXCLUSIVE OR. These are defined by their truth tables in Table II. With this method, the function F of Table I can be expressed as $F = \bar{X}\bar{Y}Z + \bar{X}Y\bar{Z} + XY\bar{Z}$. This form of expression is known as the sum-of-products form, and each term is called a minterm. The number of terms can be minimized by applying logical reduction techniques based on the theorems of Boolean algebra. For example, the above expression can be simplified to $F = \bar{X}\bar{Y}Z + Y\bar{Z}$.

TABLE II Basic Logic Operations

NOT

X	$\bar{X}$
0	1
1	0

AND

X	Y	XY
0	0	0
0	1	0
1	0	0
1	1	1

OR

X	Y	$X+Y$
0	0	0
0	1	1
1	0	1
1	1	1

NAND

X	Y	$\overline{XY}$
0	0	1
0	1	1
1	0	1
1	1	0

NOR

X	Y	$\overline{X+Y}$
0	0	1
0	1	0
1	0	0
1	1	0

EXCLUSIVE OR

X	Y	$X \oplus Y$
0	0	0
0	1	1
1	0	1
1	1	0

B. Relevant System Theory

1. Linear Shift-Invariant (LSI) Systems

A system is defined as a mapping of an input function $\mathbf{f}(x)$ into an output function $\mathbf{g}(x)$. This is represented as $\mathbf{g}(x) = S\{\mathbf{f}(x)\}$, where $S\{\mathbf{f}(x)\}$ indicates a general function operating on $\mathbf{f}(x)$. Both the input and the output are functions of x, which could be time, distance, or another parameter. A system is linear if it satisfies the superposition property, defined as follows:

$$S\{\mathbf{k}_1\mathbf{f}_1(x) + \mathbf{k}_2\mathbf{f}_2(x)\} = \mathbf{k}_1 S\{\mathbf{f}_1(x)\} + \mathbf{k}_2 S\{\mathbf{f}_2(x)\}, \quad (18)$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ are two arbitrary input functions and $\mathbf{k}_1$ and $\mathbf{k}_2$ are two arbitrary complex constants.

A system is called shift-invariant if a shift in the input results only in an equal shift in the output as follows:

$$S\{\mathbf{f}(x - x_0)\} = \mathbf{g}(x - x_0). \quad (19)$$

If the input and output are functions of time, this property is known as time-invariance. In optical systems, the input and output are usually two-dimensional spatial functions, so this property is known as space-invariance.

2. Properties of LSI Systems

Linear shift-invariant systems have interesting properties that make their analysis much simpler than shift-variant systems. One of the important features of an LSI system is that if the response of the system to an impulse function is known, the output of the system can be predicted for any input. This is achieved through the convolution process. The output is given by the following:

$$\mathbf{g}(x) = \int_{-\infty}^{+\infty} \mathbf{f}(\alpha)\mathbf{h}(x - \alpha)\, d\alpha, \quad (20)$$

where $\mathbf{h}$ represents the impulse response of the system. The analysis of LSI systems usually becomes easier if the Fourier transforms of the functions are used. The above convolution is then replaced by the following simple relation:

$$\mathbf{G}(u) = \mathbf{F}(u)\mathbf{H}(u), \quad (21)$$

where $\mathbf{F}(u)$, $\mathbf{G}(u)$, and $\mathbf{H}(u)$ are the Fourier transforms of the input function, output function, and impulse response, respectively. $\mathbf{H}(u)$ is known as the transfer function. The characteristics of an LSI system are completely specified by its transfer function.

C. Relevant Optics

There are five commonly used representations of optics: geometrical optics, wave optics, Fourier optics, quantum optics, and relativistic optics. Although these treatments are on the same general subject, each is suitable for analyzing a particular group of phenomena. To the lowest order

of approximation, images obtained by lenses or mirrors are analyzed by the laws of geometrical optics. Aperture diffraction, grating diffraction, and scattering by particles are among the topics that require consideration of the wave nature of light. Fourier optics is a useful tool for studying optical systems. Some phenomena, such as the photoelectric effect, can be explained only by quantum optics, and finally, relativistic optics is based on Einstein's postulates and can explain modern experimental measurements related to the speed of light. Geometrical optics, wave optics, and Fourier optics are all useful in analyzing optical information processing systems. Some aspects of the wave optics and Fourier optics are explained in this section.

1. Basic Concepts

In wave optics, light is treated as an electromagnetic wave. If the light source radiates at a fixed frequency, the light is called monochromatic. If the light is not monochromatic but has a relatively narrow spectrum, it is known as quasimonochromatic. Light is called polarized if the direction of the electric field is predictable (e.g., the output of a laser) and unpolarized if the orientation of the electric field is random (e.g., the radiation from an incandescent lamp). The polarization could be linear, circular, or elliptical. If the direction of the electric field at a point in space is always along the same line, the light is linearly polarized. If the two components of the electric field along the x and y axes have equal magnitude but differ in phase by 90°, the tip of the electric field vector rotates on a circle and the light is circularly polarized. Elliptical polarization corresponds to the case where the two components of the field are not equal and/or they differ in phase by an arbitrary angle. Actually, elliptical polarization is a general case, and linear and circular polarizations can be considered as special cases of it.

A linearly polarized monochromatic wave can be represented by a scalar function $u(x, y, z; t)$ that can be a component of the electric field or the magnetic field. This scalar function can be represented as follows:

$$u(x, y, z; t) = U(x, y, z)\cos[2\pi f t + \phi(x, y, z)] \quad (22)$$

where $U(x, y, z)$ and $\phi(x, y, z)$ are the amplitude and phase of the wave, respectively, and f is the frequency of the light. With complex notation, this equation can be written as follows:

$$\begin{aligned} u(x, y, z; t) &= \mathrm{Re}[\mathbf{u}(x, y, z; t)] \\ &= \mathrm{Re}[\mathbf{U}(x, y, z)\ \exp(j2\pi f t)], \end{aligned} \quad (23)$$

where Re means the real part, $\mathbf{U}(x, y, z) = U(x, y, z) \times \exp[j\phi(x, y, z)]$ is a complex amplitude, and $\mathbf{u}(x, y, z; t) = \mathbf{U}(x, y, z)\exp(j2\pi f t)$. It is common to drop the time-dependent exponential term and work with the complex amplitude. However, the physical field at any point can be obtained only after the complex amplitude is multiplied by $\exp(j2\pi f t)$ and the real part of the result is taken. The above complex amplitude should satisfy the time-independent wave equation, known as the Helmholtz equation.

Two particular waves of interest are the plane wave and the spherical wave. The names indicate the shapes of the surfaces of constant phase (wavefronts). The complex amplitude of a plane wave can be written as follows:

$$\mathbf{U}(x, y, z) = A\ \exp[jk(\gamma_x x + \gamma_y y + \gamma_z z)], \quad (24)$$

where A is the wave amplitude and $k = 2\pi/\lambda$ where λ is the wavelength. The parameters γ_x, γ_y, and γ_z are the cosines of the angles between the direction of the field propagation and the x, y, and z axes, respectively. If a spherical wave is generated by a source at $P_0 = (x_0, y_0, z_0)$, the complex amplitude corresponding to a point $P_1 = (x_1, y_1, z_1)$ is given by the following:

$$\mathbf{U}(x, y, z) = \frac{A}{r_{01}}\exp(jkr_{01}), \quad (25)$$

where A is a constant and $r_{01} = [(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2]^{1/2}$ is the distance between points P_1 and P_0. If the source and the observation points are restricted to regions near the z axis as follows:

$$(z_1 - z_0)^3 \gg \frac{\pi}{4\lambda}\left[(x_1 - x_0)^2 + (y_1 - y_0)^2\right]^2 \quad (26)$$

then Eq. (25) can be approximated as the following:

$$\begin{aligned} \mathbf{U}(x, y, z) \simeq\ & \frac{A}{z_1 - z_0}\exp[jk(z_1 - z_0)] \\ & \times \exp\left[\frac{jk\left[(x_1 - x_0)^2 + (y_1 - y_0)^2\right]}{2(z_1 - z_0)}\right]. \end{aligned} \quad (27)$$

This is known as the quadratic approximation of a spherical wave.

In general, radiation is a statistical phenomenon, and light should be described by probability functions. The information about the statistics of light is usually provided by coherene functions. The coherence of a wave field is a measure of its ability to produce observable constructive and destructive interference (i.e., bright and dark regions) when different portions of it are combined. A distinction is often made between temporal coherence and spatial coherence. These are mathematically described by two complex functions. The temporal coherence function $\mathbf{\Gamma}$ is defined as follows:

$$\mathbf{\Gamma}(\tau) = \langle \mathbf{u}(x, y; t)\mathbf{u}^*(x, y; t - \tau)\rangle, \quad (28)$$

where $\mathbf{u}(x, y; t)$ and $\mathbf{u}(x, y; t-\tau)$ are the analytic functions corresponding to the electric field intensity at point (x, y) and at instants t and $t-\tau$. The angle brackets represent an infinite time average. For a monochromatic light, $\mathbf{u}$ has a single frequency and $\mathbf{\Gamma}$ has a magnitude that is independent of τ. Such a completely temporal coherent light does not exist. In practice, a coherent light source has a band of frequencies. As a result, the magnitude of the coherent function decreases as τ is increased. This can be experimentally observed as a decrease in the visibility of the interference pattern as the time delay between the two portions of the beam is increased. The coherence time of a light source is defined as the value of τ for which the magnitude of coherence function falls to $1/e$ of its maximum value. Lasers usually have coherence times in the order of milliseconds in comparison with picoseconds coherence times of typical spectral sources. For time intervals greater than the coherence time, the source is considered to be incoherent. Another useful parameter is the complex degree of coherence, $\gamma(\tau)$. This is the normalized form of the temporal coherence function which is obtained by dividing $\mathbf{\Gamma}(\tau)$ by its maximum value [i.e., $\mathbf{\Gamma}(0)$]. A monochromatic light has a degree of coherence with a magnitude of unity, while a completely incoherent light has a degree of coherence of zero.

The parallel definition for spatial coherence function, $\mathbf{\Gamma}(\eta, \xi)$, is as follows:

$$\mathbf{\Gamma}(\eta, \xi) = \langle \mathbf{u}(x, y; t)\, \mathbf{u}^*(x-\eta, y-\xi; t) \rangle. \tag{29}$$

In this case, the analytic function $\mathbf{u}$ is measured at two points that are located at a plane perpendicular to the light rays. Again, for monochromatic light, $\mathbf{\Gamma}$ has a constant magnitude. In practice, a light source has a coherence length, which is defined as the spatial distance at which the magnitude of $\mathbf{\Gamma}$ falls to $1/e$ of its maximum value at $\eta = \xi = 0$.

Another parameter of interest in wave optics is the light intensity, which is defined as the radiated power per unit area. The intensity of monochromatic light is proportional to the square of its electric field amplitude. Neglecting a constant factor that depends on the unit system used, the following equation:

$$I(x, y) = |\mathbf{U}(x, y)|^2 \tag{30}$$

relates the complex amplitude $\mathbf{U}$ to the intensity I of monochromatic light. In general, the intensity is defined as the following:

$$I(x, y, z) = 2\langle u^2(x, y, z; t) \rangle, \tag{31}$$

where the angle brackets represent an infinite time average.

Inside a material, light travels slower than in a vacuum. The ratio of the speed of light in a vacuum to the speed of light in a material is called the index of refraction of that material. Some materials, known as photorefractive materials, have the property that their index of refraction can be altered by exposing them to light. The mechanism of the photorefractive effect is as follows. When the material is exposed to light, extra charged particles (electrons and holes) are created. These particles then migrate in certain directions and produce regions of excess positive and negative charge inside the material. The electric field induced by this spatial charge distribution changes the index of refraction of the material.

2. Diffraction Phenomena

To the lowest degree of approximation, the spreading of light when it encounters obstacles or apertures is described by ray-tracing rules from geometrical optics. The wave nature of light is ignored in this analysis, and the result becomes inaccurate as the dimensions of the obstacle or aperture are decreased. This departure from the predictions of geometrical optics is representative of diffraction phenomena.

There are two common treatments of diffraction: scalar diffraction theory and vectorial diffraction theory. In the former method, light is treated as a scalar parameter that represents one component of the electric field or magnetic field, and it is assumed that the other components behave similarly. This simplifies the analysis, but it ignores the fact that the components of the electric and magnetic fields are coupled by Maxwell's equations. This coupling is considered in the vectorial theory, but the relations obtained, except for few cases, are too complicated to solve. Fortunately, the results from the scalar theory are very accurate if two conditions are satisfied. These are (1) $\lambda/d \ll 1$, where λ is the light wavelength and d the maximum linear dimension of the aperture or obstacle; and (2) $r \gg d$, where r is the distance at which the diffracted light is observed. Both conditions are satisfied in most optical systems of interest here.

A case of great interest in optics is the diffraction by a finite planar aperture (Σ) shown in Fig. 4. In this figure, the vector $\mathbf{n}$ is an outward vector of unit amplitude that is normal to the plane of aperture. It is assumed that the incident wave is a monochromatic plane wave traveling along the z axis and the diffracted light is observed at a distance z. Assuming that the complex amplitude corresponding to a component of the incident electric field is known, one seeks the corresponding component at the observation plane. The scalar diffraction theory for this problem was first developed by Kirchhoff in 1882. In his analysis, Kirchhoff made two assumptions about the values of the complex amplitude ($\mathbf{U}$) and its normal derivative ($\partial\mathbf{U}/\partial\mathbf{n}$) immediately behind the $z = 0$ plane. These assumptions,

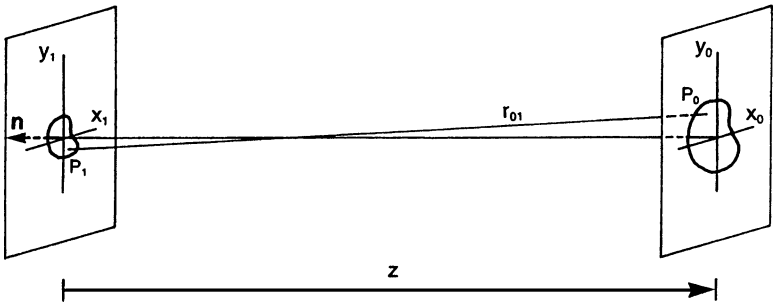


FIGURE 4 Diffraction geometry, where the vector **n** is of unit length and is normal to the aperture plane.

known as Kirchhoff's assumptions, are (1) for the points on the aperture [i.e., $(x_1, y_1) \in \Sigma$], $\mathbf{U}$ and $\partial\mathbf{U}/\partial\mathbf{n}$ are the same as they would be in the absence of the screen and (2) for the points outside the aperture [i.e., $(x_1, y_1) \notin \Sigma$], $\mathbf{U}$ and $\partial\mathbf{U}/\partial\mathbf{n}$ are zero. Based on these assumptions, Kirchhoff found the following relation for the field at P_0:

$$\mathbf{U}(P_0) = \frac{1}{4\pi} \iint_{\Sigma} \frac{\exp(jkr_{01})}{r_{01}} \times \left[\frac{\partial \mathbf{U}}{\partial \mathbf{n}} - jk\mathbf{U}\cos(\mathbf{n}, \mathbf{r}_{01}) \right] ds, \quad (32)$$

where $r_{01} = [(x_0 - x_1)^2 + (y_0 - y_1)^2 + z^2]^{1/2}$ is the magnitude of the vector $\mathbf{r}_{01}$ that joins point P_0 at the observation region to point P_1 on the aperture and $(\mathbf{n}, \mathbf{r}_{01})$ is the angle between $\mathbf{n}$ and $\mathbf{r}_{01}$.

Kirchhoff's assumptions were later shown to be inconsistent. Sommerfeld used another approach to remove this inconsistency. However, the results obtained from Kirchhoff's formula are in remarkably good agreement with experiments, and his theory is still widely used. Sommerfeld's formula for the same system is as follows:

$$\mathbf{U}(P_0) = \frac{1}{j\lambda} \iint_{\Sigma} \mathbf{U}(P_1) \frac{\exp(jkr_{01})}{r_{01}} \cos(\mathbf{n}, \mathbf{r}_{01})\, ds. \quad (33)$$

This equation is simplified if it is assumed that the distance between the two planes is much larger than both the dimension of the aperture and the dimension of the region of interest in the observation plane. Based on these assumptions, the cosine term of Eq. (33) can be approximated by unity, and the r_{01} term in the denominator can be replaced by z. A further approximation can be made in the exponent term if z is large enough to satisfy the following condition:

$$z^3 \gg (\pi/4\lambda)\left[(x_0 - x_1)^2 + (y_0 - y_1)^2\right]^2_{\max}. \quad (34)$$

Under these conditions, Eq. (33) can be approximated as follows:

$$\mathbf{U}(x_0, y_0) = \frac{\exp(jkz)}{j\lambda z} \int_{-\infty}^{+\infty}\int \mathbf{U}(x_1, y_1) \times \exp\left\{ j\frac{k}{2z}\left[(x_0 - x_1)^2 + (y_1 - y_0)^2\right] \right\} dx_1\, dy_1, \quad (35)$$

where the limits of the integral have been changed assuming $\mathbf{U}(x_1, y_1) = 0$ outside the aperture. This formula represents Fresnel diffraction, and the region in which it is valid is referred to as the Fresnel region.

If the distance between the planes satisfies the more restrictive condition as follows:

$$z \gg (k/2)\left(x_1^2 + y_1^2\right)_{\max}, \quad (36)$$

Eq. (35) can be written as follows:

$$\mathbf{U}(x_0, y_0) = \frac{\exp(jkz)}{j\lambda z} \exp\left[\frac{jk\left(x_0^2 + y_0^2\right)}{2z} \right] \int_{-\infty}^{+\infty}\int \mathbf{U}(x_1, y_1) \times \exp\left[\frac{-jk(x_0x_1 + y_0y_1)}{z} \right] dx_1\, dy_1. \quad (37)$$

This represents Fraunhofer diffraction, and the region that satisfies condition (36) is called the Fraunhofer region. Because the terms in front of the integral are constants, this formula can be simplified to the following:

$$\mathbf{U}(x_0, y_0) = C \int_{-\infty}^{+\infty}\int \mathbf{U}(x_1, y_1) \times \exp\left[\frac{-j2\pi(x_0x_1 + y_0y_1)}{\lambda z} \right] dx_1\, dy_1, \quad (38)$$

where $C = [\exp(jkz)/j\lambda z] \exp[jk(x_0^2 + y_0^2)/2z]$. Equation (38) can be considered as a two-dimensional Fourier transform of the function $\mathbf{U}(x_1, y_1)$ evaluated at spatial

frequencies $f_X = x_0/\lambda z$ and $f_Y = y_0/\lambda z$. This result indicates that the shape of the diffraction pattern does not change in the Fraunhofer region. The change in the z value enters only as a scaling factor.

If the aperture is covered by a transparency having a complex transmittance of $\mathbf{t}(x_1, y_1)$, the above formulas are valid provided that $\mathbf{U}(x_1, y_1)$ is replaced by $\mathbf{U}(x_1, y_1) = \mathbf{t}(x_1, y_1)\,\mathbf{U}^-(x_1, y_1)$, where $\mathbf{U}^-(x_1, y_1)$ is the complex amplitude of the incident wave just before reaching the aperture. For the case of an incident plane wave $\mathbf{U}^-(x_1, y_1) = A$, where A is the wave amplitude. The Fraunhofer diffraction then becomes the Fourier transform of $\mathbf{t}(x_1, y_1)$ multiplied by a constant.

3. Fourier Transform Properties of Lenses

Lenses are among the most important components of the optical information processing systems. A lens is composed of an optically dense and transparent material like glass. A lens is called thin if the output light emerges at approximately the same coordinates as the input light. Figure 5 shows a lens and its parameters. The effect of a lens on the incident light can be considered as producing a phase factor delay of the following:

$$\mathbf{t}(x, y) = \exp(jkn\Delta_0)\exp\left[\frac{-jk(x^2 + y^2)}{2f}\right], \qquad (39)$$

which, when multiplied by the complex amplitude of the field just in front of the lens ($\mathbf{U}_1$), provides the complex amplitude of the field just behind the lens ($\mathbf{U}_1'$). The parameter f in Eq. (39) is the focal length of the lens, which is given by the following:

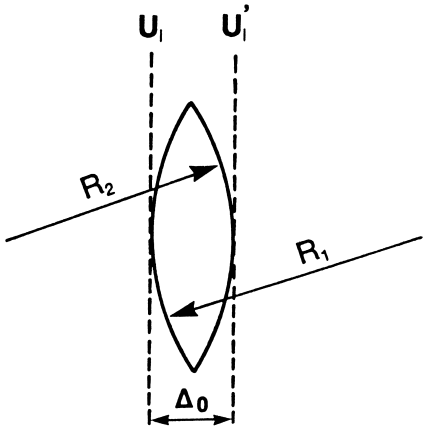


FIGURE 5 Thin lens, where R_1 and R_2 are the radii of the lens surfaces; $\mathbf{U}_1$ is the complex amplitude of the input wave; $\mathbf{U}_1'$ the complex amplitude of the output wave; and Δ_0 the maximum thickness of the lens.

$$f = \frac{1}{(n-1)[(1/R_1) - (1/R_2)]}. \qquad (40)$$

This formula is valid for all kinds of lenses if the following convention is used to determine the signs of R_1 and R_2: As rays travel from left to right, the radius of each convex surface encountered is considered positive, and the radius of each concave surface is considered negative. Therefore, R_1 in Fig. 5 is positive, while R_2 is negative. Finally, n is the index of refraction.

Practically, Eq. (39) is not valid for all values of x and y, since the lens has a finite aperture. The effective transmittance function is obtained by multiplying $\mathbf{t}(x, y)$, as given in Eq. (39), by a pupil function $P(x, y)$. The pupil function is unity for a point on the lens aperture (Σ) and is zero for a point outside the aperture as follows:

$$P(x, y) = \begin{cases} 1, & (x, y) \in \Sigma \\ 0, & (x, y) \notin \Sigma. \end{cases} \qquad (41)$$

An important property of a lens is its capability of Fourier transforming. It can be shown that if an object with a complex transmittance of $\mathbf{t}(x, y)$ is placed at a distance $d_1 (d_1 \geq 0)$ in front of a lens or at a distance $d_2 (f > d_2 \geq 0)$ behind the lens, and an incident monochromatic plane wave illuminates the object, the Fourier transform of $\mathbf{t}(x, y)$ multiplied by a constant and by a phase factor is obtained at the back focal plane. This property can be described as the following:

$$\mathbf{U}_\mathrm{f}(x_\mathrm{f}, y_\mathrm{f}) = (A/j\lambda f)C\mathcal{F}\{\mathbf{t}(x, y)\}\big|_{\substack{f_X = x_\mathrm{f}/\lambda f \\ f_Y = y_\mathrm{f}/\lambda f}}, \qquad (42)$$

where x_f and y_f are the coordinates of a point in the focal plane, $\mathbf{U}_\mathrm{f}(x_\mathrm{f}, y_\mathrm{f})$ is the complex amplitude of the wave at the back focal plane, A the amplitude of the incident wave, and C a phase factor. The constant $j\lambda f$ is usually ignored in the analysis. For the special case that the object is at the front focal plane ($d_1 = f$), the phase factor is unity and an exact Fourier transform is obtained.

Two types of lenses are widely used in optical information processing: spherical lenses and cylindrical lenses. The surface of a spherical lens has curvature in both directions. This type can perform a two-dimensional Fourier transform as discussed above. A cylindrical lens has curvature in one dimension (Fig. 6); hence, it can perform a one-dimensional Fourier transform. The cylindrical lenses are used for multichannel processing where different data channels are stacked along the axis of the lens.

4. Analysis of Optical Imaging Systems

A simple imaging system that consists of a single positive lens is shown in Fig. 7. From geometrical optics, if the following relation:

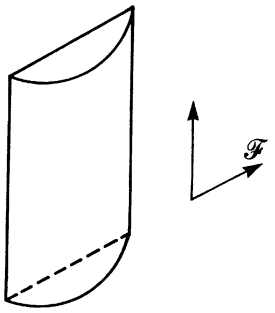


FIGURE 6 Cylindrical lens, where the direction along which the Fourier transform is performed is indicated by $\mathcal{F}$.

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f} \tag{43}$$

is satisfied, an image of linear magnification $m = d_i/d_o$ is obtained. A more rigorous treatment includes the effects of diffraction of light between the two planes. In this analysis, it is assumed that the incident light is monochromatic. The finite aperture of the lens or the physical aperture usually used with the lens is modeled by a pupil function $P(x, y)$, defined in Eq. (41). The result obtained from the diffraction analysis shows that the complex amplitude of the image, $\mathbf{U}_i(x_i, y_i)$, can be obtained from the complex amplitude of the image predicted by the geometrical optics, $\mathbf{U}_g(x_i, y_i)$, using the following:

$$\mathbf{U}_i(x_i, y_i) = C\tilde{\mathbf{h}}(x_i, y_i) * \mathbf{U}_g(x_i, y_i). \tag{44}$$

The parameter C is a quadratic phase factor that does not affect the intensity of the image, and it is normally ignored. The function $\tilde{\mathbf{h}}(x_i, y_i)$ is related to the pupil function through

$$\tilde{\mathbf{h}}(x_i, y_i) = \left(\frac{1}{\lambda d_i}\right)^2 \hat{P}\left(\frac{x_i}{\lambda d_i}, \frac{y_i}{\lambda d_i}\right), \tag{45}$$

where $\hat{P}(x_i/\lambda d_i,\ y_i/\lambda d_i)$ represents the Fourier transform of the pupil function evaluated at the spatial frequencies $f_X = x_i/\lambda d_i$ and $f_Y = y_i/\lambda d_i$.

In general, imaging systems are more complex. They usually include several lenses and apertures. Also, pure monochromatic light is not available. The effects of all apertures in the system can be modeled by one effective aperture located behind the last lens. The pupil function corresponding to this effective aperture can be obtained by projecting all the apertures to the location of the last lens and choosing the common area. The quasi-monochromaticity of the light is measured by the coherence function. In the analysis of imaging systems, illuminations with high degree of coherence are considered coherent, and those with low degree of coherence are considered incoherent.

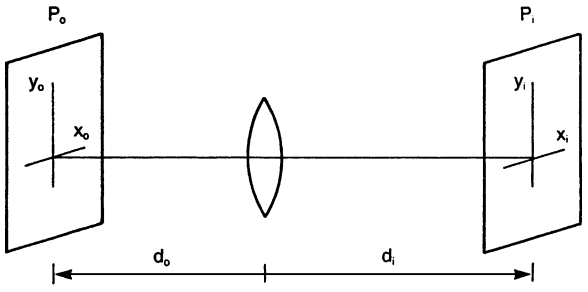


FIGURE 7 Simple imaging system, where d_i is the image distance; d_o, object distance; P_i, image plane; and P_o, object plane.

In a coherently illuminated system, since the phase relation between the different waves that reach a point on the image plane is fixed, the waves add together as complex amplitudes. Therefore, the results obtained for the monochromatic case are valid for coherent systems, with the understanding that $\mathbf{U}$ now represents the relative amplitude and phase of the light. The analysis of coherent systems becomes simpler if performed in the spatial frequency domain. The convolution Eq. (44) can then be replaced by the following:

$$\hat{\mathbf{U}}_i(f_X, f_Y) = \mathbf{H}(f_X, f_Y)\,\hat{\mathbf{U}}_g(f_X, f_Y), \tag{46}$$

where $\hat{\mathbf{U}}_i$ and $\hat{\mathbf{U}}_g$ represent the Fourier transforms of $\hat{\mathbf{U}}_i$ and $\hat{\mathbf{U}}_g$, respectively, and $\mathbf{H}(f_X, f_Y) = \mathcal{F}\{\tilde{\mathbf{h}}(x_i, y_i)\}$ is called the coherent transfer function of the system. It can be shown that the transfer function is related to the effective pupil function by the following:

$$\mathbf{H}(f_X, f_Y) = P(-\lambda d_i f_X, -\lambda d_i f_Y); \tag{47}$$

that is, the transfer function is a magnified and inverted version of the effective pupil.

In a noncoherent illuminated system, the waves that reach a point on the image plane are statistically independent. Therefore, they add on a power (or intensity) basis. As a result, the system is linear in intensity, and the image intensity $I_i(x_i, y_i)$ can be obtained from the following:

$$I_i(x_i, y_i) = C|\tilde{\mathbf{h}}(x_i, y_i)|^2 * I_g(x_i, y_i), \tag{48}$$

where C is a constant, $I_g(x_i, y_i)$ the intensity of the image obtained from geometrical optics and $\tilde{\mathbf{h}}(x_i, y_i)$ is defined in Eq. (45). The equivalent form of the above equation in the frequency domain is as follows:

$$\hat{I}_i(f_X, f_Y) = C\mathcal{F}\left\{|\tilde{\mathbf{h}}(x_i, y_i)|^2\right\}\hat{I}_g(f_X, f_Y). \tag{49}$$

It is common to normalize the parameters in Eq. (49) to their values at $f_X = f_Y = 0$. The normalized form of $\mathcal{F}\{|\tilde{\mathbf{h}}(x_i, y_i)|^2\}$ is called optical transfer function (OTF) and is represented by $\mathcal{H}(f_X, f_Y)$. OTF has a maximum value of unity at the origin.

III. COHERENT OPTICAL PROCESSING

A. Spectrum Analysis

As shown in Section II.C.3, the Fourier transform of the transmittance of a coherently illuminated object can be obtained by a simple lens. While the Fourier transform is a computationally intensive operation with an electronic digital computer, it is performed in parallel with optics and thus is very fast. The input to an optical spectrum analyzer is a planar transducer with an amplitude transmittance $\mathbf{t}(x, y)$ that represents the function under the investigation. If the Fourier transform of a complex function is desired, the input transducer should be capable of modulating both the amplitude and phase of the light. This can be achieved by a two-layer photographic film. In the case of real functions, a simple black-and-white film suffices. The spectrum obtained at the back focal plane can be detected visually or electronically, or it can be recorded on another photographic film. Depending on the system configuration, as described below, the Fourier transform of two-dimensional or one-dimensional signals can be obtained.

A characteristic parameter of an optical spectrum analyzer is the space-bandwidth product. This is analogous to the time-bandwidth product of an electronic amplifier. The space-bandwidth product of an optical system is defined as the product of the highest spatial frequency over which it can operate and the size of the input aperture. Comparing two optical systems, the system with the larger space-bandwidth product is capable of analyzing higher spatial frequencies for a given aperture size or larger aperture sizes for a given spatial frequency. Space-bandwidth products of up to 10^5 are achievable with high-precision optical systems. Even with an ordinary 35-mm camera lens, space-bandwidth products greater than 1000 can be obtained.

The only shortcoming of optical spectrum analyzers is their low accuracy. The accuracy of these systems depends on the energy level of the light source, the quality of the optical components, and the noise of the detecting system. The limiting factor is normally the noise of the photodetector. The accuracies of practical systems are typically in the range of 1 to 10%. This limits the practical application of these systems to areas where high speed is required while low accuracy is tolerable.

1. Two-Dimensional Spectrum Analysis

A schematic diagram of a two-dimensional spectrum analyzer is shown in Fig. 8. A beam expander system provides plane-wave illumination from the output of the laser. The system consists of two spherical lenses and a pinhole aperture. The first lens, L_1, acts as a condenser. At the front focal plane of L_1, a pinhole aperture with a diameter of about 10 μm is used to filter out the undesired modes and to provide a point source of light to the second lens, L_2. The second lens, located at a focal length from the pinhole, collimates the light and provides a uniform illumination to the input plane. The input and output planes are located at the front and back focal planes of the third lens, L_3, to satisfy the condition for an exact Fourier transform. In most practical systems, L_3 is replaced by an optical system that consists of several lenses to improve the quality of the transform. The frequency components of the input function can be obtained by measuring the coordinates of the bright spots at P_2 and using the relations $f_X = x/\lambda f_3$

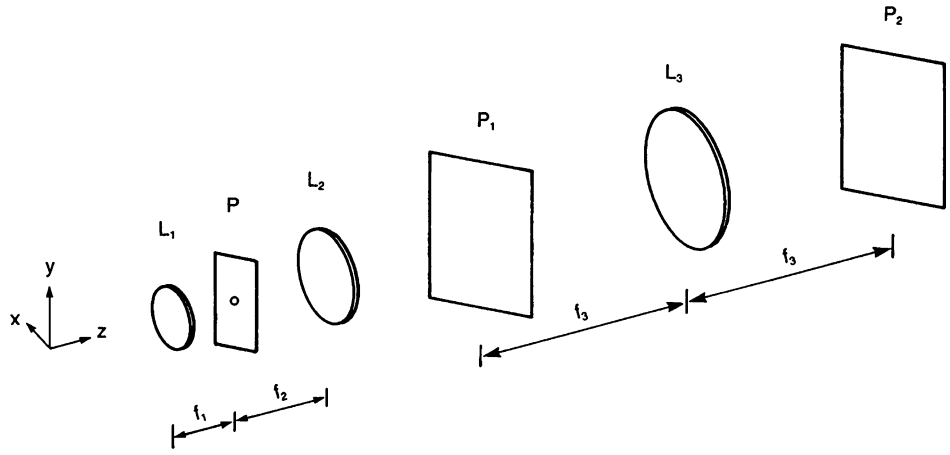


FIGURE 8 Schematic diagram of a two-dimensional spectrum analyzer, where L_1 is the condensing lens; L_2, collimating lens; L_3, Fourier transform lens; f_1, f_2, and f_3 are the focal lengths of lenses L_1, L_2, and L_3, respectively; P is the pinhole aperture; P_1, input plane; and P_2, output plane.

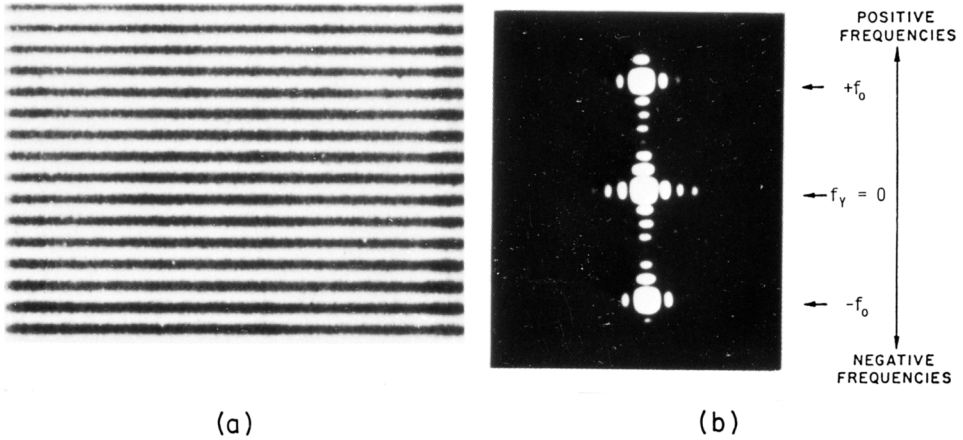


FIGURE 9 Example of (a) input and (b) output of an optical spectrum analyzer. [From Preston, K., Jr. (1972). "Coherent Optical Computers," McGraw–Hill, New York.]

and $f_Y = y/\lambda f_3$, where λ is the free-space wavelength of the light. Figure 9 shows an example of the input and the output of the spectrum analyzer as recorded on photographic films. The input is a biased one-dimensional sinusoidal function as follows:

$$\mathbf{t}(x, y) = b + a \sin(2\pi f_o y), \tag{50}$$

where $b > a$ to ensure a positive real function. The Fourier transform of this function consists of three delta functions as follows:

$$\mathcal{F}\{\mathbf{t}(x, y)\} = b\,\delta(f_X, f_Y) - j\frac{a}{2}\,\delta(f_X, f_Y - f_o) + j\frac{a}{2}\,\delta(f_X, f_Y + f_o). \tag{51}$$

These three impulses appear as three bright spots along the f_Y axis in Fig. 9b. The extra spots visible in that figure are due to the finite dimension of the input aperture. The Fourier transform of a rectangular aperture is a two-dimensional sinc function. Since the input function is multiplied by a rectangle function at P_1, the spectrum obtained at P_2 is the convolution of Eq. (51) with a sinc function.

2. Multichannel Spectrum Analysis

The schematic diagram of a multichannel spectrum analyzer is shown in Fig. 10. This is similar to Fig. 8 except that the spherical lens, L_3, is replaced by three cylindrical lenses (L_3, L_4, and L_5). The input consists of several channels stacked along the vertical axis (y). The amplitude transmittance of each channel represents a one-dimensional function. The combination of the three cylindrical lenses performs the Fourier transformation along

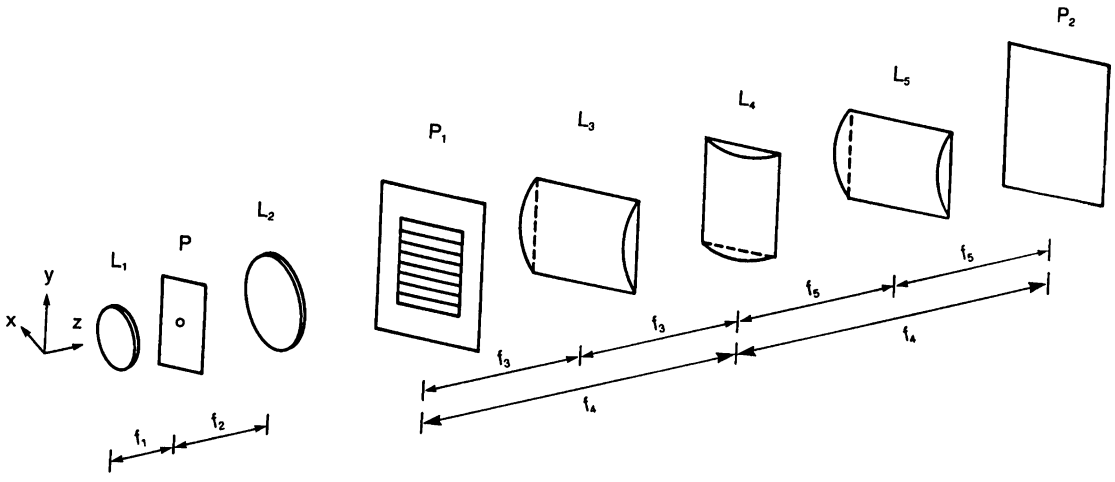


FIGURE 10 Schematic diagram of a multichannel one-dimensional spectrum analyzer, where f_3, f_4, and f_5 are the focal lengths of cylindrical lenses L_3, L_4, and L_5, respectively. Other parameters are the same as in Fig. 8.

the x axis and images along the y axis. As a result, the Fourier transform of each channel is obtained in a horizontal band located at a particular distance from the x axis. To prevent the spectra of the neighboring channels from affecting each other, spatial guard bands should be considered in the input plane. Assuming 20 channels/mm and using an active region of 27 mm in 35-mm diameter optics, one can obtain 540 channels.

3. Folded-Spectrum Technique

As described in the previous section, the spectrum analysis of one-dimensional functions can be performed by multichannel systems. It is sometimes desired to analyze the spectrum of a long one-dimensional signal. This can be done by dividing the signal into a number of segments and providing those segments as the inputs to a multichannel system. In the output plane, the spectra obtained should be integrated at each frequency over all channels. Such an integration can be implemented by a vertical slit and a photomultiplier. The photomultiplier integrates the light that is passed through the slit, and the entire assembly moves along the x axis to detect all the frequencies. The problem with this system is its low space-bandwidth product. Actually, the multichannel capability has been sacrificed to analyze a long signal, but since each segment is treated individually, the space–bandwidth product remains the same. Therefore, the resolution of the system is not increased. To increase the space–bandwidth product, a method was proposed and analyzed by Thomas in 1966 and later implemented and demonstrated by Markevitch in 1969. Thomas suggested that a two-dimensional spectrum analyzer, as shown in Fig. 8, be used for analyzing long signals. The input signal is recorded in a raster format, but the different segments are not isolated as different channels. In fact, since a two-dimensional Fourier transformation is performed, all the segments contribute to the output spectrum in both directions. Thomas showed that the space–bandwidth obtained by this method is the square of the value that is achieved by the multichannel technique. Some of the properties of the spectrum obtained are described here. For a more detailed analysis, the reader is referred to the original article by Thomas [Thomas, C. E. (1966). *Applied Optics* **5**(11), 1782–1790].

Figure 11a shows the raster scan input format. The input signal, which is a time-dependent function, is used to modulate the intensity of an electron beam as it scans the screen. The scan rate is adjusted so that the maximum frequency of the input signal generates a resolvable spatial frequency on the film. This is achieved using a deflection speed of $V_s = f_{max}/p$, where f_{max} is the maximum frequency of the signal and p the number of resolvable line pairs per unit length of the film. The exposed film is

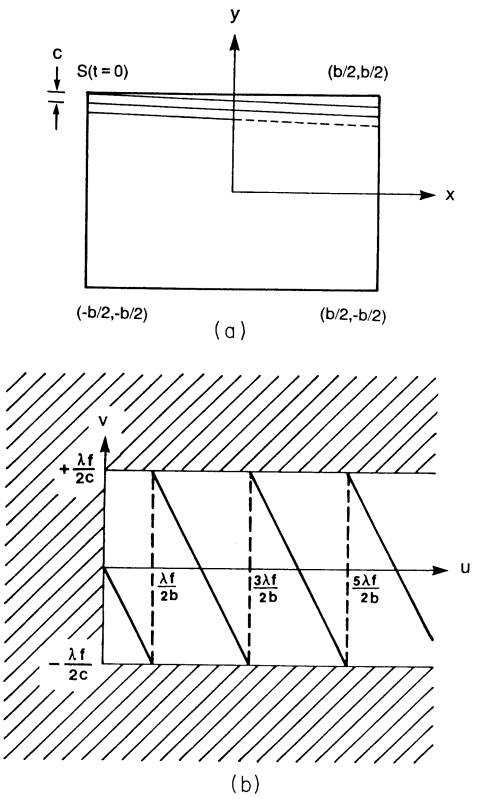


FIGURE 11 Raster scan (a) input and (b) output formats in folded spectrum technique, where f is the focal length of the Fourier transforming lens; S, recorded signal; and λ, free-space wavelength of the laser. [From Thomas, C. E. (1966). *Applied Optics* **5**(11), 1782–1790.]

then appropriately linearized during the developing process. Therefore, the transmittance of the developed film along the scan lines resembles the amplitude of the signal. The spectrum of the recorded signal has a periodic structure along the v axis that is due to the effective periodic sampling nature of the raster lines. The information about the spectrum can be obtained from one of these periodic regions, such as the one shown in Fig. 11b. Since the amplitude spectrum of a real signal is a symmetric function, the analysis of only positive frequencies is sufficient. If the recorded signal is a sinusoidal wave, its frequency appears as a bright spot in the above region. If the spectra of waveforms with different frequencies are obtained, it can be seen that as the frequency is increased, the bright spot on the output plane scans through the observed region. These output scan lines are tilted with respect to the input lines by 90°.

Thomas has shown that the number of resolution elements along each line of the output spectrum is equal to the number of input scan lines. Conversely, the number of scan lines in the output spectral plane is equal to the number of resolution elements along each scan line of the input

data. Thus, the obtained spectrum has as many resolution elements as the input data. Due to the nature of the appearance of this spectrum, it is known as the folded spectrum. Also, the spatial frequency axes u and v are referred to as the coarse frequency and fine frequency axes, respectively. Each raster line in the output plane represents a part of the spectrum with a bandwidth of $\Delta f = 1/T$, where T is the total length of the recorded signal in seconds.

The power of the folded spectrum technique can be seen by way of a numerical example. Using a photographic film with a resolution of 50 line pairs/mm, one can record 200 msec of a 10-MHz bandwidth signal on a 2×2-cm area. The corresponding folded spectrum consists of 1000 lines with 2000 resolvable spots on each line. The total space–bandwidth product of the system is 2×10^6, and the system is capable of detecting frequencies up to 10 MHz with a resolution of 5 Hz.

B. Spatial Filter Synthesis

1. System Configuration

In spatial filtering techniques, the Fourier transform of an input function that is obtained by a lens is manipulated by a filter. A second lens performs the Fourier transform operation on the modified spectrum and provides the output.

Figure 12 shows an optical system that is commonly used for spatial filtering analysis. The input plane P_1 is illuminated by a plane wave that propagates along the z axis. the function $\mathbf{f}(x, y)$ is introduced at P_1 as the amplitude transmittance of either a photographic film or a spatial light modulator device. An exact Fourier transform of the input function is obtained at the focal plane, P_2, as $\mathbf{F}(f_X, f_Y)$, where the spatial frequencies f_X and f_Y are related to the coordinates x_2 and y_2 through the relations $f_X = x_2/\lambda f$ and $f_Y = y_2/\lambda f$. As light passes through the filter plane, the Fourier transform is multiplied by the amplitude transmittance function of this plane. A second Fourier transformation is then performed on the filtered spectrum by L_3. Since two succeeding Fourier transforms result in an inversion of the input function, the axes of the output plane P_3 are usually inverted to simplify the analysis. It is not essential that L_2 and L_3 have the same focal length. However, this is usually assumed, and the system is referred to as the 4-f system. Also, the constants in Eq. (42) are usually ignored. Considering these points, one can write the output function at P_3 as follows:

$$\mathbf{g}(x, y) = \mathbf{f}(x, y) * \mathbf{h}(x, y), \tag{52}$$

where $\mathbf{h}(x, y)$ is the inverse Fourier transform of the function introduced at P_2. In the frequency domain, the equivalent relation to the above equation is the following:

$$\mathbf{G}(f_X, f_Y) = \mathbf{F}(f_X, f_Y)\mathbf{H}(f_X, f_Y), \tag{53}$$

where $\mathbf{G}$, $\mathbf{F}$, and $\mathbf{H}$ are the Fourier transforms of $\mathbf{g}$, $\mathbf{f}$, and $\mathbf{h}$, respectively.

2. Simple Filters

Simple filters are filters that alter either the amplitude or the phase of the spectrum. An amplitude filter can be obtained by the process of developing an exposed photographic film. A phase filter requires the control of optical path length of the filter plate, for example, by an appropriately varying thickness. A special type of amplitude filter is the binary filter which consists of opaque and transparent regions.

Because of their construction simplicity, these filters were among the first implemented filters. Porter's experiment, described in Section I, is an example of binary filtering. Zernike's phase-contrast microscope was based on introducing an additional phase of 90° to the undiffracted components of light. A useful operation in image processing is to attenuate the low-frequency components by an amplitude filter. Since the average brightness of the image is determined by these components, this filtering results in decreasing the average brightness and increasing the contrast.

3. Holographic Filters

In spatial filtering, it is usually necessary to find the Fourier transform of a desired impulse response and to construct the corresponding filter plate. The Fourier transform of

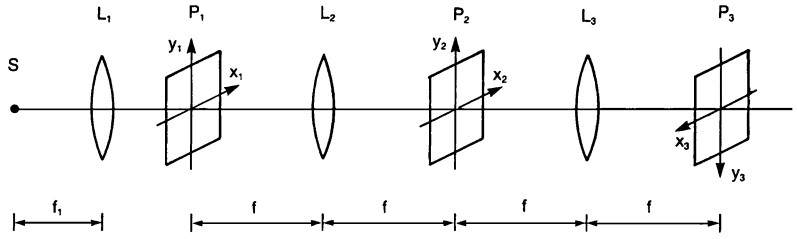


FIGURE 12 Configuration of a spatial filtering system where f_1 is the focal length of lens L_1; f, focal length of lenses L_2 and L_3; L_1, collimating lens; L_2 and L_3 are Fourier transforming lenses; P_1 is the object plane; P_2, focal plane; P_3, output plane; and S, light source.

an arbitrary two-dimensional function is computationally intensive, and the result is usually a complex function. Although complex filters can be constructed by cascading an amplitude filter and a phase filter, practical problems in fabricating phase filters make the task difficult. To overcome these difficulties, Vander Lugt introduced holographic filters in 1963.

Figure 13a shows the recording process for a Vander Lugt filter. The desired impulse response $\mathbf{h}(x, y)$ is introduced as the amplitude transmittance of the input transparency. The Fourier transform of the impulse response $\mathbf{H}(f_X, f_Y)$ is obtained optically at the back focal plane, P_2, of a lens, where it is interfered with a tilted plane wave. The resultant interference pattern is recorded on a photographic film. Although the developed film consists solely of absorption patterns, it contains the information about both amplitude and phase of the desired transfer function. This can be seen from the intensity pattern at P_2. The tilted wavefront can be represented by $A\exp(-j2\pi y_2 \sin\theta/\lambda)$, where A is the amplitude and θ the angle of the tilted wave. At P_2 this complex amplitude is added to $(1/j\lambda f)\mathbf{H}(x_2/\lambda f, y_2/\lambda f)$ which is obtained by L_1. The resultant intensity, given by the following:

$$I(x, y) = |A\exp(-j2\pi y_2 \sin\theta/\lambda) + (1/j\lambda f)\mathbf{H}(x_2/\lambda f, y_2/\lambda f)|^2, \qquad (54)$$

is recorded as a hologram. The recorded hologram is then placed at the filter plane, P_2, in the system shown in Fig. 12. The input function $\mathbf{f}(x, y)$ is introduced at P_1,

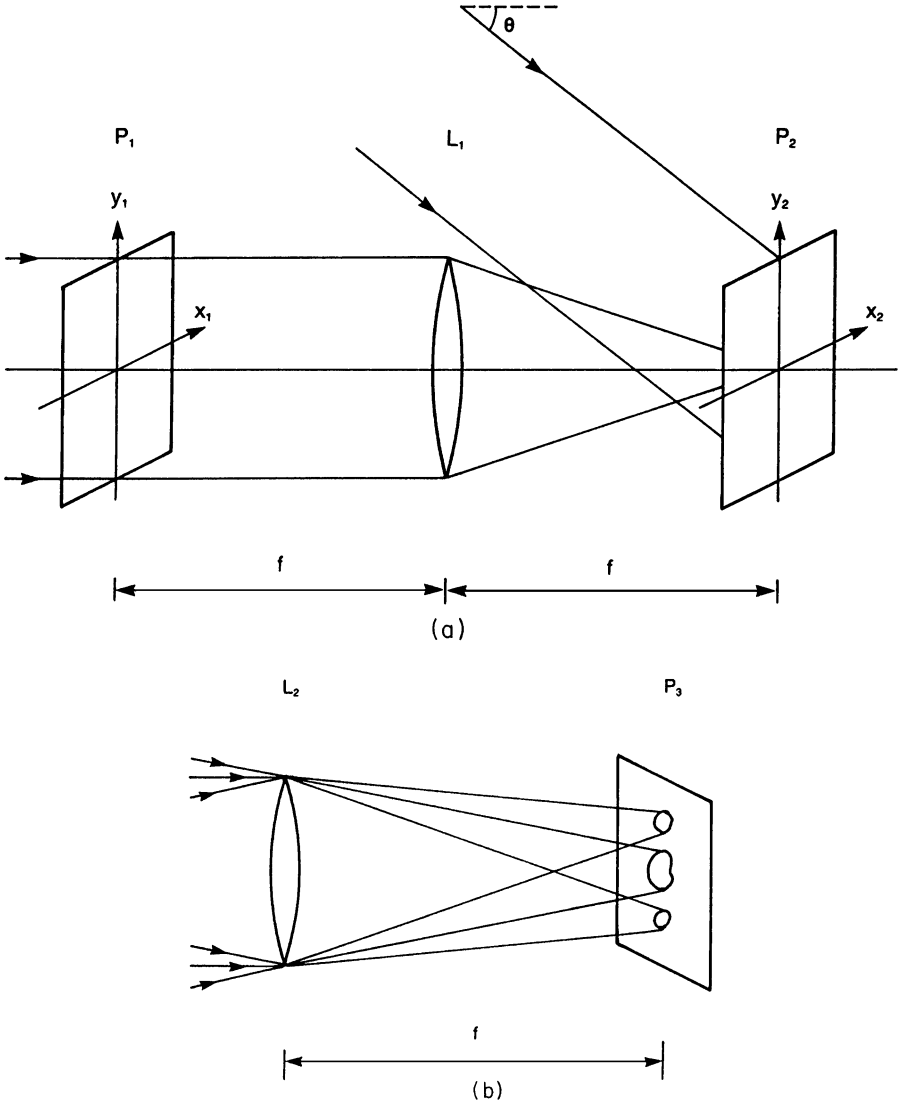


FIGURE 13 Vander Lugt filter: (a) recording process and (b) output, where f is the focal length of lenses L_1 and L_2; L_1 and L_2 are Fourier transforming lenses; P_1 is the input plane; P_2, focal plane; and P_3, output plane.

and its Fourier transform is obtained at P_2. At P_2, the light is diffracted by the hologram. This is equivalent to multiplying the Fourier transform of $\mathbf{f}(x, y)$ by the transmittance function of the hologram, which is the same as the above intensity pattern. The diffracted light has four components. Two terms propagate along the z axis and are of no interest. The other two terms propagate at angles $+\theta$ and $-\theta$ with respect to the z axis and provide the convolution and cross-correlation of the input function with the impulse response. This is shown in Fig. 13b. To prevent overlapping of different terms, the following condition:

$$\theta > (W_f + 1.5\, W_\mathrm{h})/f \tag{55}$$

should be satisfied at the recording process, where W_f and W_h are the maximum dimensions of the functions $\mathbf{f}(x, y)$ and $\mathbf{h}(x, y)$ along the y direction.

The Vander Lugt filter has several advantages. It is capable of implementing complex filters with a single amplitude filter. The Fourier transform of the impulse response is obtained optically, and finally, both convolution and cross-correlation results are provided in the output.

4. Computer-Generated Filters

Although the Vander Lugt filter simplifies the implementation of complex spatial filters, it requires the physical existence of the impulse response. In many cases of interest, the mathematical description of the desired impulse response is known, but it is difficult or even impossible to construct it physically. These cases can be handled using a digital computer. Computer-generated holograms were first introduced by Brown and Lohmann in 1966. One of the advantages of their method was that complex spatial filtering was performed with binary masks. These masks were obtained by photographic reduction of the output of a computer-controlled plotter. Other techniques for making computer-generated holograms have been introduced since Brown and Lohmann's invention. Today, computer-generated holograms are widely used in various applications and are considered to possess outstanding potential for future optical information-processing systems. Because of space constraints, only a brief description of Brown and Lohmann's method is presented here.

The holograms introduced by Brown and Lohmann are known as detour-phase holograms. The name indicates the method used to achieve the required phase variations. It is well known that a periodic array of slits (a grating) diffracts an incident wave in different directions. These directions can be calculated using the following relation:

$$(\sin\alpha_n - \sin\alpha_0)\, d = n\lambda, \tag{56}$$

where n is an integer that indicates the order of the diffracted wave, α_0 the angle of the incident wave, α_n the angle of the nth diffracted beam, d the spacing between the slits, and λ the free space wavelength of the light. If some of the slits are shifted from their periodic positions, the phase of the diffracted light is affected as shown in Fig. 14. This is known as the detour phase and can be used to implement the phase of a complex function.

To make a computer-generated hologram, the complex function that describes the transmittance of the desired filter is sampled in both the x and y directions according to the sampling theorem. That is, the sampling period is chosen to be less than or equal to $1/2f_\mathrm{max}$, where f_max is the maximum spatial frequency along the direction of sampling. At each sampled point, the value of the function is represented by an amplitude and a phase, such as $A_{nm}\exp(j\phi_{nm})$, where n and m are the indices of the sampled point along the x and y directions, respectively. A computer-controlled plotter is used to draw an enlarged version of the filter. This plot consists of rectangular cells corresponding to the sampled values. Each cell, as shown in Fig. 15, includes a transparent aperture on an opaque background. The complex value of the wavefront is encoded by the size and position of the aperture. One method is to choose the height of the aperture proportional to the

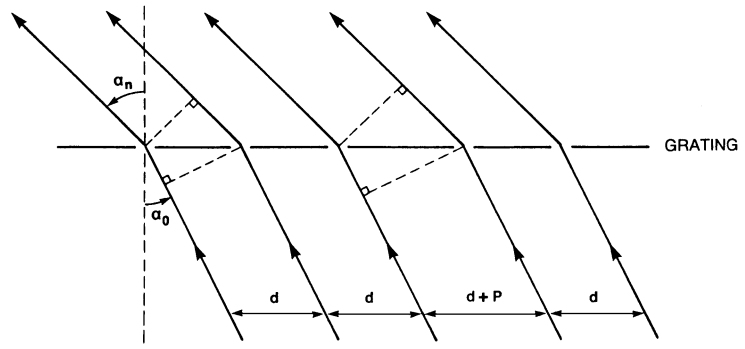


FIGURE 14 Detour phase effect, where α_0 is the angle of the incident wave and α_n, angle of the nth diffracted beam. [From Lohmann, A. W., and Paris, D. P. (1968). *Applied Optics* **7**(4), 651–655.]

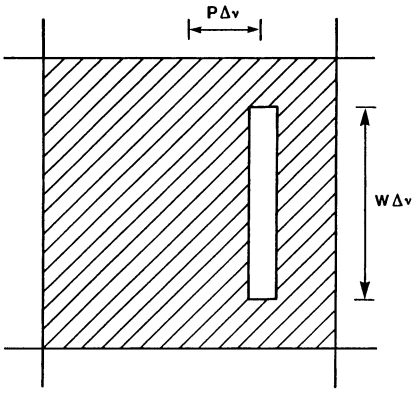


FIGURE 15 Example of a sampling cell in a detour phase hologram, where Δv is the resolution of the plotter and W and P are proportional to the amplitude and phase of the sampled wavefront.

amplitude and its position proportional to the phase. This plot is then photographically reduced to the actual size of the filter, which is usually about 1–2 cm. The size of the apertures after reduction is typically around 10–20 μm. Because of the binary nature of this filter, it can be easily reproduced. Also, film grain noise and film nonlinearities are of little consequence. Computer-generated holograms are made as arrays of at least 128 × 128 pixels.

5. Other System Configurations

The system shown in Fig. 12 is not the only processor used for spatial filtering. Figure 16a shows a two-lens processor that is widely used. The distances in this system are such that (1) the point source, S, is imaged at plane P_2 by lens L_1 and (2) the input plane P_1 is imaged at the output plane P_3 by lens L_2. It can be shown that a Fourier transform of an input function at P_1 is obtained at P_2 where it is affected by the filter. One of the advantages of this system is that the scale of the Fourier transform can be varied by changing the distance between P_1 and P_2. Also, the distances between P_1, L_2, and P_3 can be selected to obtain the desired magnification. Figure 16b shows a one-lens system that can be used for spatial filtering. In this system, the point source, S, is imaged at plane P_2 by lens L_1, and the input plane P_1 is imaged at the output plane P_3 by the same lens.

Another useful system is the joint transform correlator shown in Fig. 17. This is a 4-f system with two input functions, $\mathbf{g}_1(x, y)$ and $\mathbf{g}_2(x, y)$, that are vertically separated in the input plane by a distance Δ. First, a Fourier hologram is recorded at P_2. Given a linear recording of the light intensity, the transmittance of the resultant hologram is as follows:

$$\begin{aligned}\mathbf{t}(f_X, f_Y) &= |\mathbf{G}_1(f_X, f_Y)\exp(j\pi\Delta f_Y) \\ &\quad + \mathbf{G}_2(f_X, f_Y)\exp(-j\pi\Delta f_Y)|^2 \\ &= |\mathbf{G}_1(f_X, f_Y)|^2 + |\mathbf{G}_2(f_X, f_Y)|^2 \\ &\quad + \mathbf{G}_1(f_X, f_Y)\,\mathbf{G}_2^*(f_X, f_Y)\exp(j2\pi\Delta f_Y) \\ &\quad + \mathbf{G}_1^*(f_X, f_Y)\,\mathbf{G}_2(f_X, f_Y)\exp(-j2\pi\Delta f_Y),\end{aligned} \tag{57}$$

where $f_X = x_2/\lambda f$ and $f_Y = y_2/\lambda f$ are the spatial frequencies and $\mathbf{G}_1$ and $\mathbf{G}_2$ the Fourier transforms of $\mathbf{g}_1$ and $\mathbf{g}_2$,

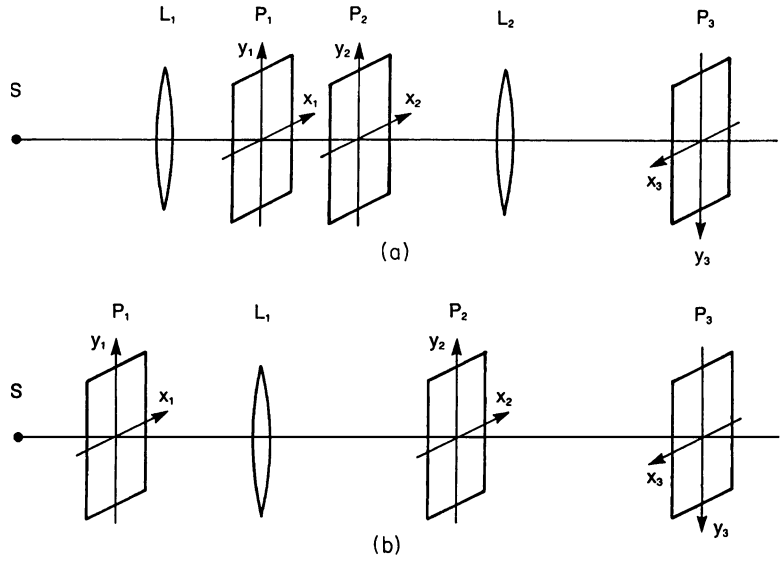


FIGURE 16 (a) Two-lens coherent optical processor. (b) One-lens coherent optical processor. L_1 and L_2 are lenses; P_1 is the input plane; P_2, filter plane; P_3, output plane; and S, light source.

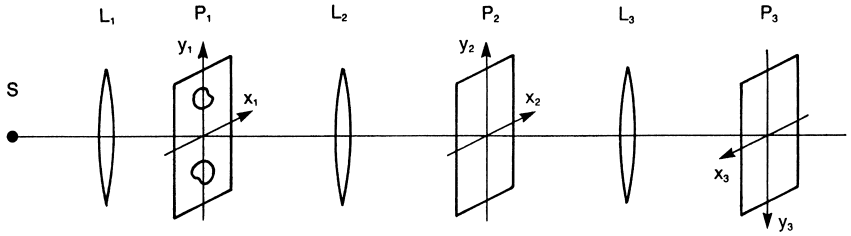


FIGURE 17 Joint transform correlator, where L_1 is the collimating lens; L_2 and L_3, Fourier transform lenses; P_1, input plane; P_2, focal plane; P_3, output plane; and S, light source.

respectively. Then, the input plane is removed, and the Fourier transform of the above function is obtained at P_3. The third and fourth terms of the transmittance expression produce the cross-correlation of the two functions around the $(x_3 = 0, y_3 = +\Delta)$ and $(x_3 = 0, y_3 = -\Delta)$ points, respectively. To separate these two terms from the other two terms that appear near the origin, the following relation:

$$\Delta \geq 0.5(W_1 + W_2) + \max(W_1, W_2) \tag{58}$$

should be satisfied, where W_1 and W_2 are the dimensions of the two functions along the y axis.

IV. INCOHERENT OPTICAL PROCESSING

A. Introduction

Coherent optical processing suffers from several problems. Data from commonly used noncoherent sources, such as LEDs and CRTs, cannot be directly processed. Dust particles on the optical components or small defects inside the lenses can produce scattered waves and affect the output significantly. Incoherent systems do not have these problems. To compare the two types of processing, the corresponding systems are shown in Fig. 18. In a coherent system, the entire object is illuminated by a plane wave. This is similar to a one-channel system. If a dust particle exists on one of the lenses, a particular region of the output is affected. In the case of an incoherent system, the light source is extended and can be considered as a number of point sources. Each point of the object plane, the Fourier plane, or the output plane is illuminated by waves that originate from different points of the source. This is similar to a multichannel system. As a result, a dust particle on a lens does not produce a severe effect in any particular region of the output. The effect actually spreads to all regions but becomes more tolerable since each point in the output plane is illuminated by different waves. Therefore, the overall signal-to-noise ratio is slightly decreased without having severe effects in a particular region of the output.

Despite these advantages, incoherent systems have their own shortcomings. A major problem is that they are inherently capable of handling only nonnegative real functions. This is due to the fact that they work by manipulating the light intensity, not the complex amplitude. Although, as described below, there are methods that include bipolar and complex functions, they usually require hybrid systems. Also, the total information capacity of incoherent systems, in general, falls below that of coherent systems. However, because of the unique features of incoherent optical processing mentioned above, it is the preferred choice for some applications.

Incoherent optical processors can be divided into two groups: (1) the systems that are based on the diffraction phenomenon and (2) the systems that can be explained by geometrical optics rules. These are explained in the following two subsections. Methods for including bipolar and complex functions are also discussed and the corresponding problems are addressed.

B. Diffraction-Based Systems

Figure 19 shows a diffraction-based incoherent optical processor. The object can be sel-fluminous, or it can be illuminated by an extended source. As discussed in Section II.C.4, an incoherent system is linear in intensity. The intensity of the output image, $I_i(x, y)$, can be obtained from the following:

$$I_i(x, y) = Cs(x, y) * I_g(x, y) \tag{59}$$

where C is a constant and $I_g(x, y)$ the intensity of the image that is predicted by geometrical optics. The function $s(x, y)$ represents the diffraction effects of the system. It is known as the point-spread function and is defined as the intensity distribution that is obtained in the image plane if a point source is located at the origin of the object plane. For the system of Fig. 19, the following obtains:

$$s(x, y) = \left(\frac{1}{\lambda f}\right)^4 \left|\hat{\mathbf{P}}\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right)\right|^2, \tag{60}$$

where $\hat{\mathbf{P}}(x/\lambda f, y/\lambda f)$ represents the Fourier transform of the complex pupil function (complex transmittance of the aperture) at spatial frequencies $f_X = x/\lambda f$ and

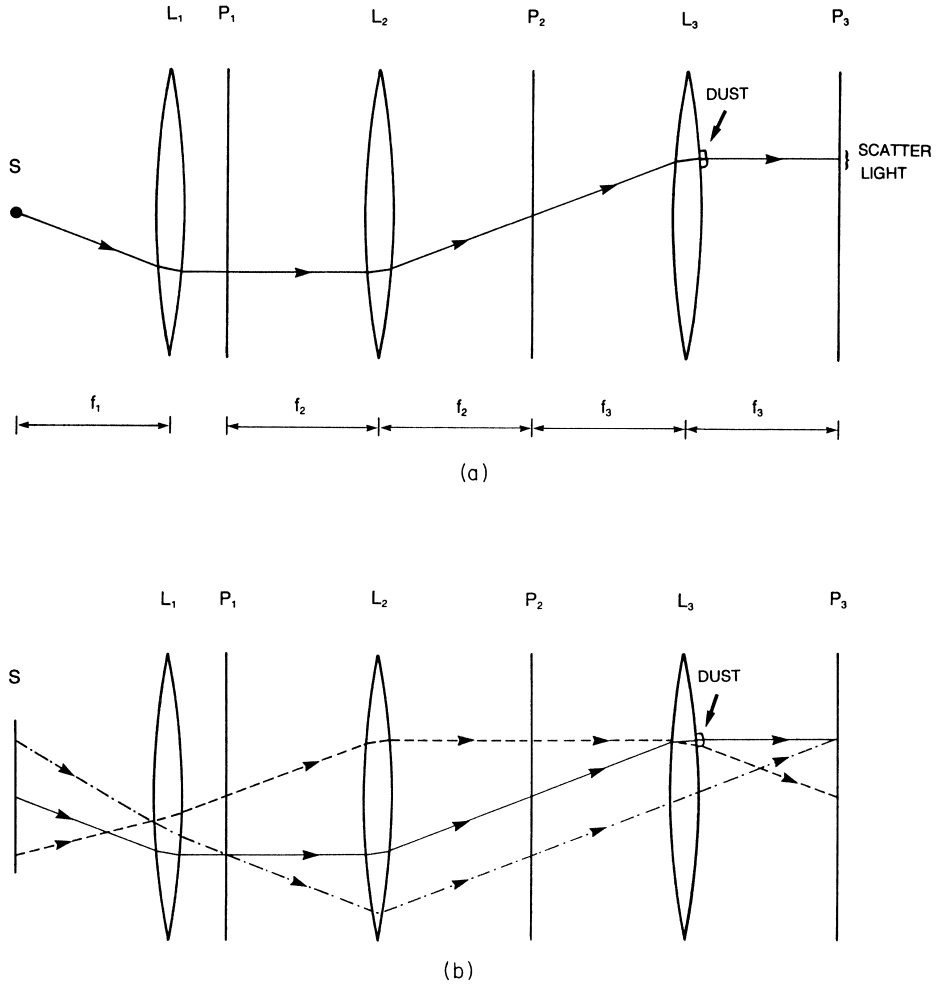


FIGURE 18 Effect of dust particles in (a) coherent and (b) incoherent optical systems, where f_1, f_2, and f_3 are the focal lengths of L_1, L_2, and L_3, respectively; L_1 is the collimating lens; L_2 and L_3 are the Fourier transform lenses; P_1 is the object plane; P_2, Fourier plane; P_3, output plane; and S, light source. [From Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.]

$f_Y = y/\lambda f$. It is convenient to use a scaled pupil function defined as $\mathbf{P}'(x, y) = \mathbf{P}(\lambda f_x, \lambda f y)$. Equation (60) is then simplified to the following:

$$s(x, y) = |\hat{\mathbf{P}}'(x, y)|^2. \tag{61}$$

Since the image predicted by geometric optics is an inverted version of the object, the following relation:

$$I_g(x, y) = I_o(-x, -y) \tag{62}$$

holds between the intensity of the object, I_o, and the intensity of the image, I_g. The axes of the image plane are usually inverted to make the two distributions identical. Equation (59) can then be written as follows:

$$I_i(x, y) = Cs(x, y) * I_o(x, y). \tag{63}$$

The equivalent form of this relation in the Fourier domain is the following:

$$\hat{I}_i(f_X, f_Y) = C\hat{s}(f_X, f_Y)\hat{I}_o(f_X, f_Y), \tag{64}$$

where $\hat{}$ represents the Fourier transform operation. Using Eq. (61), one can find the Fourier transform of the point-spread function as the autocorrelation of the scaled pupil function as follows:

$$\hat{s}(f_X, f_Y) = \mathbf{P}'(f_X, f_Y) \star \mathbf{P}'(f_X, f_Y). \tag{65}$$

The functions in Eq. (64) are usually normalized by their values at the origin. The normalized form of $\hat{s}(f_X, f_Y)$ [i.e., $\hat{s}(f_X, f_Y)/\hat{s}(0, 0)$] is known as the optical transfer function and is usually denoted by $\mathcal{H}(f_X, f_Y)$. Using the normalized quantities, one can rewrite Eq. (64) as follows:

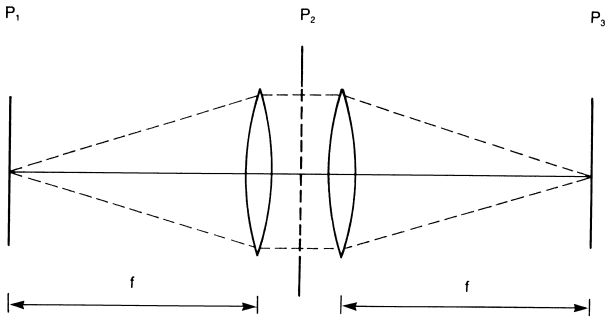


FIGURE 19 Diffraction-based incoherent optical processor, where f is the focal length of the lenses; P_1, object plane; P_2, pupil plane; and P_3, output plane. [From Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.]

$$\hat{I}_{\mathrm{i}}(f_X, f_Y) = \mathscr{H}(f_X, f_Y)\hat{I}_{\mathrm{o}}(f_X, f_Y). \tag{66}$$

This equation is the basis for the spatial filtering process using incoherent systems. It shows that the spatial-frequency components of the light intensity at the object plane can be manipulated in the output plane by an appropriate optical transfer function.

The first diffraction-based incoherent optical processor was introduced for character recognition by Armitage and Lohmann in 1965. Their system is shown in Fig. 20. The left part of the system is coherent and provides the Fourier transform of the transmittance of the input plane, $\mathbf{U}_n(x, y)$, at the back focal plane of lens L_1. The intensity distribution at P_1, $I_n(x, y)$, is related to $\mathbf{U}_n(x, y)$ by the following equation:

$$I_n(x, y) = C_1|\hat{\mathbf{U}}_n(x/\lambda f, y/\lambda f)|^2, \tag{67}$$

where C_1 is a constant. A rotating plate made of ground glass is used to make the phase a random variable. The rest of the system works as an incoherent processor. The point-spread function of the incoherent system is $s(x, y) = (1/\lambda f)^4|\hat{\mathbf{P}}_m(x/\lambda f, y/\lambda f)|^2$, where $\hat{\mathbf{P}}_m$ is the pupil function corresponding to the character of interest. With these relations, the distribution of light intensity in the output plane can be found from Eq. (63) as a convolution integral. Since in the character recognition process, both $\mathbf{U}_n(x, y)$ and $\mathbf{P}_m(x, y)$ are real quantities, their Fourier transforms are symmetric, and convolution becomes similar to the correlation process. The intensity of the output light can be written as follows:

$$I_{\mathrm{out}}(x, y) = C_2 \int_{-\infty}^{+\infty}\int \left|\hat{\mathbf{U}}_n\left(\frac{\xi}{\lambda f}, \frac{\eta}{\lambda f}\right)\right|^2 \times \left|\hat{\mathbf{P}}_m\left(\frac{\xi - x}{\lambda f}, \frac{\eta - y}{\lambda f}\right)\right|^2 d\xi\, d\eta, \tag{68}$$

where C_2 is a constant. At the output plane, a pinhole is located at the $(x = 0, y = 0)$ point. A detector behind the pinhole measures the intensity of the output light at the origin given by the following:

$$I_{\mathrm{out}}(0, 0) = C_2 \int_{-\infty}^{+\infty}\int \left|\hat{\mathbf{U}}_n\left(\frac{\xi}{\lambda f}, \frac{\eta}{\lambda f}\right)\right|^2 \times \left|\hat{\mathbf{P}}_m\left(\frac{\xi}{\lambda f}, \frac{\eta}{\lambda f}\right)\right|^2 d\xi\, d\eta. \tag{69}$$

This quantity is defined as the cross-correlation coefficient between the energy functions of the two characters, that is, $|\hat{\mathbf{U}}_n(\xi/\lambda f, \eta/\lambda f)|^2$ and $|\hat{\mathbf{P}}_m(\xi/\lambda f, \eta/\lambda f)|^2$. If the input character is the same as the reference character, the two energy functions are identical, and the cross-correlation coefficient becomes maximum, indicating a match case. The advantage of this system is that a character can be recognized regardless of a spatial shift. This is due to the fact that the system works based on the intensity functions of Fourier transforms, so a spatial shift of a function does not change the intensity of its Fourier transform. The

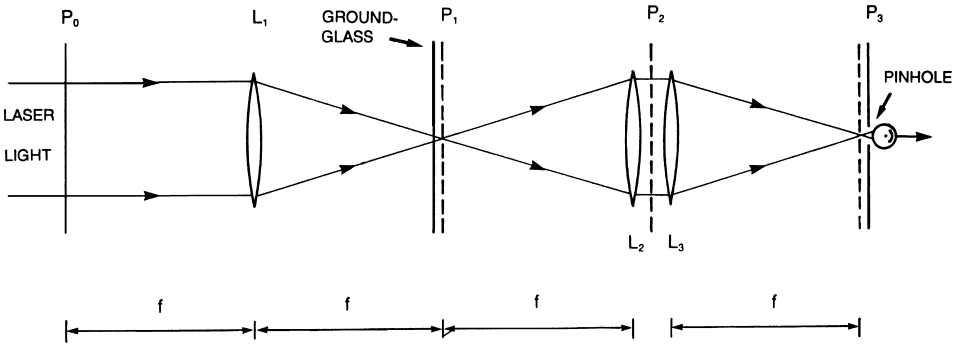


FIGURE 20 Optical system for character recognition, where f is the focal length of lenses L_1, L_2, and L_3; P_0, input plane; P_1, focal plane; P_2, pupil plane; and P_3, output plane. [From Armitage, J. D., and Lohmann, A. W. (1965). *Applied Optics* **4**(4), 461–467.]

disadvantage of the system is that some characters, such as O and Q, have similar energy functions, and therefore the corresponding correlation factors are high. As a result, the probability of an error in detection is increased for these cases.

A significant advance in the field of incoherent spatial filtering occurred in 1968 when Lohmann and, independently, Lowenthal and Werts used holographic techniques to implement the desired filters. This is the incoherent version of the Vander Lugt filtering technique. The recording process is similar to that shown in Fig. 13a, except that a diffuser is put in contact with the input transparency at P_1. As a result, the phase of the light is made a random variable. An input transparency with an intensity transmittance that resembles the desired point-spread function is used, and its Fourier hologram is recorded. This hologram is then used at the pupil plane of the system shown in Fig. 19. The results of the convolution and cross-correlation operations between the object intensity and the point spread function appear in the image plane at different locations along the y axis.

C. Geometrical Optics-Based Systems

Incoherent optical processors that are based on geometrical optics are of two types: imaging systems and shadow casting systems. In the imaging systems, the input function $f_1(x, y)$ is introduced as the intensity transmittance of a transparency at the input plane, P_1. This is imaged onto a second transparency that represents the second function $f_2(x, y)$. The output light is demagnified by a lens and spatially integrated by a detector. Several useful operations can be achieved by this scheme. A simple system is shown in Fig. 21. This system is capable of integrating the product of two functions as follows:

$$g = \int_{-\infty}^{+\infty}\int f_1(x, y) f_2(x, y)\, dx\, dy, \tag{70}$$

where g is the detected power and an imaging with unit magnification is assumed. With a spatial or temporal scanning technique, convolution and cross-correlation operations can be achieved. For example, if P_2 in Fig. 21 is shifted by x_0 and y_0 along the x and y axes, respectively, the following cross-correlation is obtained at the output:

$$g = \int_{-\infty}^{+\infty}\int f_1(x, y) f_2(x - x_0, y - y_0)\, dx\, dy. \tag{71}$$

To implement the convolution process, one of the transparencies should be inverted with respect to the coordinate axes and then shifted by the desired amount. The system can work as a two-dimensional processor or a multichannel one-dimensional processor. In the latter case, one or both of the transparencies consist of a number of stacked one-dimensional distributions. A cylindrical lens is used as L_2 to integrate only in one dimension, and the output of each channel is detected by a separate detector element.

The systems based on shadow casting can provide convolution and cross-correlation of two functions without requiring mechanical motion. Figure 22 shows a system that performs the cross-correlation process. Again, the two functions are introduced as the intensity transmittances of the transparencies at P_1 and P_2. A point at $(-x_0, -y_0)$ on the source illuminates the first transparency with a tilted plane wave. If diffraction is negligible, the tilted wave propagates between the two planes according to geometrical optics rules and illuminates the second plane. The intensity distribution of the incident light on the second plane is a shifted version of the intensity distribution on the first plane, that is, $f_1[x - (d/f)x_0, y - (d/f)y_0]$. This is multiplied by the intensity distribution of the second lens, and the result is focused to the point (x_0, y_0) on the detector. The power at this point on the detector is as follows:

$$I(x_0, y_0) = K \int_{-\infty}^{+\infty}\int f_1\left\{x - \left(\frac{d}{f}\right)x_0, y - \left(\frac{d}{f}\right)y_0\right\} \times f_2(x, y)\, dx\, dy, \tag{72}$$

where K is a constant. Therefore, the distribution of the power at the detector plane provides the cross-correlation between the two functions.

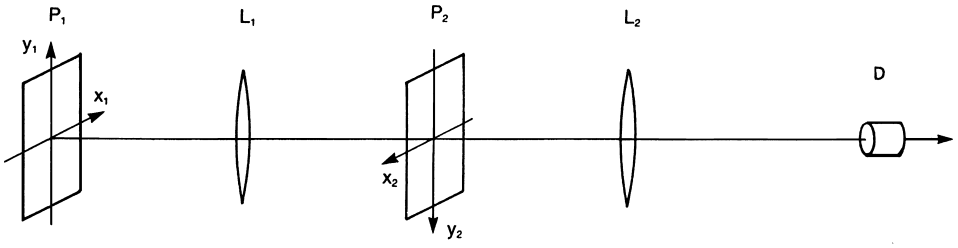


FIGURE 21 Incoherent optical system for realizing the integral of a product, where D is the detector; L_1, imaging lens; L_2, integrating lens; P_1, input plane; and P_2, image plane. [From Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.]

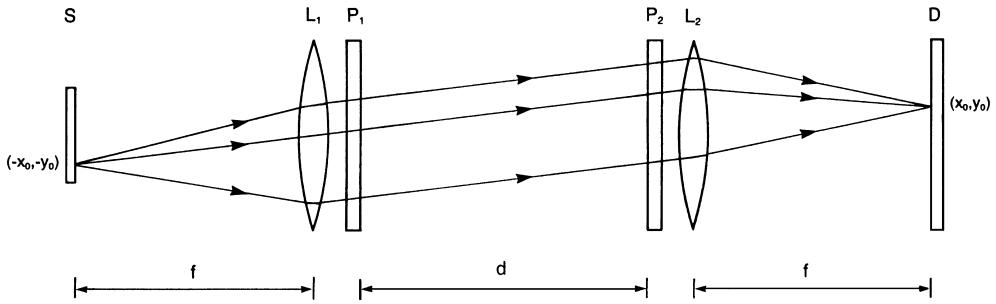


FIGURE 22 Incoherent spatial system for realizing the cross-correlation operation, where D is the detector; d, distance between the two input planes; f, focal length of lenses L_1 and L_2; P_1 and P_2 are the input planes; and S is the light source. [From Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.]

The geometrical optics-based processors are simple and powerful. However, they require that diffraction effects be negligible. This means that the transparencies should not have fine structure. As a result, the space-bandwidth products of these systems are lower than the diffraction-based systems, and in terms of information handling they are not efficient.

D. Bipolar Processing

A major limitation of incoherent optical processors is that their point-spread function is a nonnegative real quantity, while many useful operations require systems with bipolar impulse responses. A number of methods have been investigated to implement bipolar incoherent processors. The general idea behind these techniques is to use two pupil functions and get the desired operation by combining the two images at the output. One method, introduced by Rhodes, is the phase-switching technique. In this method, as shown in Fig. 23, the input light is split into two beams and passes through two spatial filters, P_1 and P_2. The beams are then recombined, and an output image is constructed. Two images are obtained with the same filters but with different phases between the beams (0° and 180°). This is achieved by moving one of the mirrors shown in Fig. 23. The two output images are subtracted from each other by an electronic system or by a digital computer.

Complex quantities can also be represented in incoherent systems. One technique is to break the real and imaginary components into positive and negative parts and use four separate pupils to manipulate them. A more efficient method, proposed by Burckhardt, is to use three unit vectors at 0°, 120°, and 240° as the basis vectors. Any complex function can then be represented by three nonnegative components.

Although these techniques expand incoherent processing beyond positive impulse response systems, they generally require hybrid systems. Consequently, the physical sizes of the systems are increased. Also, the two-pupil method does not have a unique solution for a desired point-spread function, and it is not known what the best choice is.

V. HYBRID PROCESSORS

A. Introduction

Optical processors have several important advantages. Among them are parallel processing, high throughput of data, and the ability to perform powerful operations, such as Fourier transformation, with simple systems. However, optical systems have their own shortcomings, one of which is lack of flexibility. Also, most current optical processor designs are inherently analog and cannot provide highly accurate outputs. In contrast, digital electronic systems are

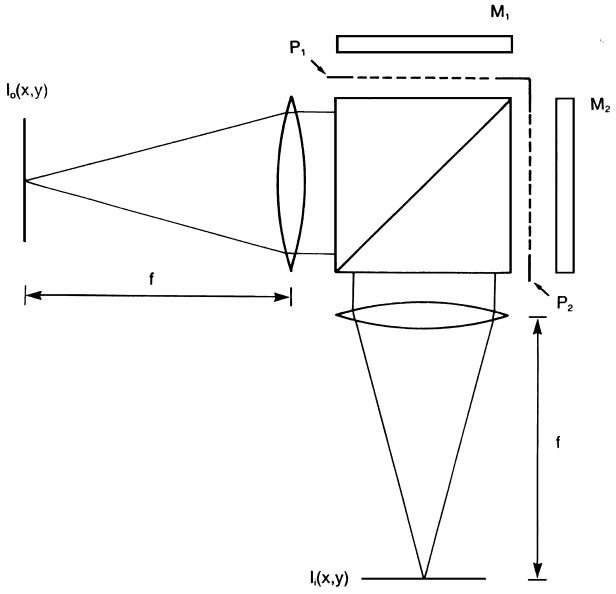


FIGURE 23 Two-pupil system based on Michelson interferometer for bipolar incoherent processing, where f is the focal length of the lenses; I_o, intensity of the input object; I_i, intensity of the output image; M_1 and M_2 are mirrors; and P_1 and P_2, pupils. [From Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.]

both flexible and accurate, and the technology of electronic computers is more advanced. Therefore, it seems reasonable to combine the two systems to utilize the advantages of both. In a hybrid (optical/electronic) processor, the optical system performs a kind of transformation on the input data. The output of the optical system is then analyzed by a digital electronic system.

One of the major design considerations of hybrid systems is the interfacing required to convert the data from one form into the other. Data are entered into the optical system through an input interface. To use the favorite features of the optical system, the input interface should have a high space-bandwidth product and be capable of handling high data rates. Two common types of input devices for incoherent optical systems are TV monitors and arrays of LEDs. Different types of spatial light modulators are used as input devices for coherent systems. The output of the optical system should be converted to an electrical signal, sampled, and digitized for use by the electronic system. These are performed by an output interface. In some hybrid systems, the spectrum of the input signal is manipulated by an interface at the Fourier plane. All three types of interfaces mentioned above are usually connected to a digital computer. Using the flexibility of computer programming, one can control the optical system and analyze the results.

Hybrid processors are made for a variety of applications, each with a special structure suitable for a particular job. A general description of the different types of hybrid systems is provided in the following subsections.

B. Coherent Hybrid Processors Based on Power Spectrum Analysis

A schematic diagram of a hybrid processor that is based on power spectrum analysis is shown in Fig. 24. A two-dimensional pattern is introduced by a spatial light modulator. The Fourier transform of this pattern is detected at the back focal plane of a spherical lens. Since the detectors are sensitive to light intensity, the detected signal is actually the square of the Fourier transform (i.e., the power spectrum). This signal is sampled and digitized. The resultant values are transmitted to an electronic digital computer. The computer uses a programmed algorithm to analyze the sampled values and to provide the desired output.

This type of processor benefits from the data compression property of Fourier transform. To get an intuitive feeling about this property, recall that the value of the Fourier transform of a function at a point in the spatial frequency domain is affected by all points in the function. If the Fourier plane is divided into a number of regions and the power spectrum is integrated inside each region, this set of numbers provides information about the function.To get detailed information, it is required to have sufficiently small regions. However, even if the regions are not small, the values provide information about the general behavior of the function, which could be sufficient for detecting the desired feature. With this technique, the amount of data is significantly reduced and becomes compatible with the capacity of the electronic system.

Different types of detectors can be used for converting the optical output to an electrical signal. A commonly used device is the wedge–ring detector. This is a two-dimensional array fabricated in silicon by planar diffusion. It has a circular shape and typically consists of 32 semicircular rings and 32 wedges; each element can be accessed individually. This structure takes advantage of the inversion symmetry property of the power spectrum. The rings provide information about the amplitudes of spatial frequency components of the spectrum, and the wedges provide information about the orientation of the spatial frequencies. If the output of each segment is represented by a

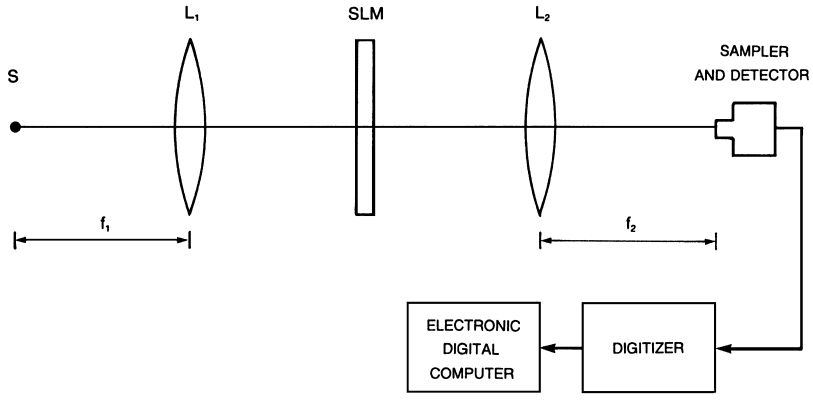


FIGURE 24 Coherent hybrid processor based on power spectrum analysis, where f_1 and f_2 are the focal lengths of lenses L_1 and L_2, respectively; L_1 is the collimating lens; L_2, Fourier transforming lens; S, light source; and SLM, spatial light modulator. [From Stark, Henry, ed. (1982). "Applications of Optical Fourier Transforms," Academic, New York.]

10-bit number, a total of 640 bits is sufficient to represent all the values. This is significantly lower than the number of bits in the input pattern. If the above detection is too coarse, other detectors such as a TV camera can be used.

Hybrid processors based on power spectrum analysis have found application areas both in the laboratory and in industry. Among the laboratory applications are analysis of aerial imagery, detection of black lung disease from chest X-ray images, and automatic sorting of handwriting. In industry, hybrid systems are used for inspecting products such as medical syringe needlepoints, woven clothes, and printing papers.

C. Coherent Hybrid Processors Based on Spatial Filtering

A schematic diagram of a hybrid processor that is based on spatial filtering is shown in Fig. 25. The optical system is a 4-f configuration as described in Section III.B.1. The electronic digital computer controls the optical system through three interfaces at the input, output, and Fourier transform planes. The input and output interfaces are similar to those described for power spectrum-based processors. The type of the third interface depends on the application of the hybrid system. If it is desired to perform different spatial filtering operations (such as low-pass or high-pass filtering), electronically controlled two-dimensional spatial light modulators are suitable. One example of such devices consists of two layers of liquid crystal. Each layer is divided into a number of regions. The regions can be activated individually by addressing the corresponding electrodes. The electrodes in one layer have ring structure and in the other layer are wedge-shaped. The ring-shaped electrodes are used to block particular spatial frequency bands, while the wedge-shaped electrodes are used to eliminate different orientations. To perform a filtering operation, the required electrodes are activated by an electronic computer. The regions in contact with these electrodes become opaque, and consequently the corresponding frequencies are blocked.

In pattern recognition, it is sometimes required to get the correlation between an input pattern and a number of reference patterns. These reference patterns could be different Vander Lugt filters recorded in a two-dimensional array format. Since the result of the correlation is very sensitive to the position of the Vander Lugt filter, an interfacing system is usually designed to move the desired filter to the required position with sufficient accuracy. These systems may be implemented with stepping motors that are controlled by digital electronic computers. The user specifies the desired filter through the computer terminal. The number of steps is determined by a program, and the specified filter is moved to within a few micrometers of the required position.

Hybrid systems have recently been used for recording high-quality computer-generated holograms. For this purpose, the Fourier transform of the desired function is first calculated and then sampled. During the recording process, the laser beam is deflected to the desired point on the film, and an exposure determined by the sampled value at that point is performed. The process is controlled by a computer for all the sampled values.

Hybrid processors based on spatial filtering are used in numerous applications. Most of these processors function

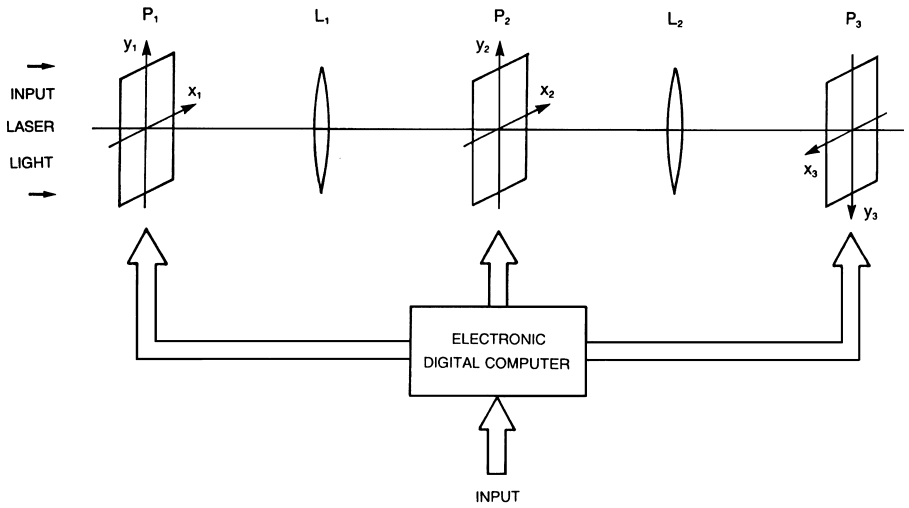


FIGURE 25 Schematic diagram of a coherent hybrid processor based on spatial filtering, where L_1 and L_2 are the Fourier transform lenses; P_1, input plane; P_2, Fourier transform plane; and P_3, output plane. [From Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.]

as correlators; they compare the input pattern(s) with one or more reference patterns and detect the matched cases. Among the applications of optical correlators are character recognition, cloud motion analysis, water pollution monitoring, radar processing, and image processing.

D. Incoherent Hybrid Processors

Incoherent processing like coherent processing benefits from the flexibility and accuracy of hybrid systems. In addition to these advantages, incoherent processing can be extended beyond nonnegative real functions by using hybrid systems. As described above, optical incoherent processors are restricted to nonnegative quantities. Bipolar and complex quantities should be decomposed into a number of nonnegative components (e.g., the positive and negative parts in the case of a bipolar quantity) and then processed separately. The components are recombined at the output to produce the desired result. These tasks can be performed by electronic systems.

As an example of incoherent hybrid processors, the system introduced by Goodman is described here (Fig. 26). This system is capable of performing vector–matrix multiplication in parallel. The values of the input vector are introduced in parallel as intensities of the LEDs. Lenses L_1 and L_2 image the light distribution horizontally and expand it vertically so that all parts of the mask M are illuminated. Lens L_3 is a field lens. The mask M is a two-dimensional array of apertures. The area of the aperture in the mth row and nth column is proportional to the value of the corresponding element in the matrix. As light passes through the mask, different product terms are obtained. The lens combination L_5 integrates the light distribution behind the mask along the horizontal direction and provides the results to a column array of detectors. Lens L_4 is a cylindrical lenslet array that improves the light throughput.

Goodman used this system to perform discrete Fourier transformations (DFTs). DFT is the analogous form of the integral Fourier transform for discrete signals. In the case of one-dimensional signal, it is defined as the following:

$$\mathbf{G}(n) = \sum_{m=0}^{N-1} \mathbf{g}(m) \exp\left(\frac{-j2\pi mn}{N}\right), \quad n = 0, 1, \ldots, N-1, \quad (73)$$

where $\mathbf{g}(m)$ is a discrete function that consists of N values and $\mathbf{G}(n)$ the nth value of the corresponding DFT. Equation (73) is actually a vector–matrix multiplication, where the vector $\mathbf{g}(m)$ is multiplied by an $N \times N$ matrix whose elements are $\exp(-j\pi 2mn/N)$. To include complex quantities, Goodman represented each complex number by three positive components, as described in Section IV.D. This requires an increase in the dimension of the system by a factor of three in both the horizontal and vertical directions.

The intensities of the LEDs in the above processor are controlled by an electronic system. Also, the detected signal is an electrical signal that can be used for another operation or can be digitized and stored. A fixed mask can be used if a particular operation is needed, or a spatial light modulator controlled by a digital computer can be used if a variety of operations is desired. This system is very powerful. It can perform various operations including space-variant and nonlinear processing. Also, since the input data are introduced in parallel, very high throughput processing can be achieved.

VI. OPTICAL DIGITAL COMPUTING

A. Introduction

The optical processors described in previous sections are all analog in nature. Although hybrid systems benefit from digital computation, they actually convert an optical analog signal to an electrical digital signal and then perform digital computation in the electronic system. To utilize the full advantage of the speed and parallelism of optics,

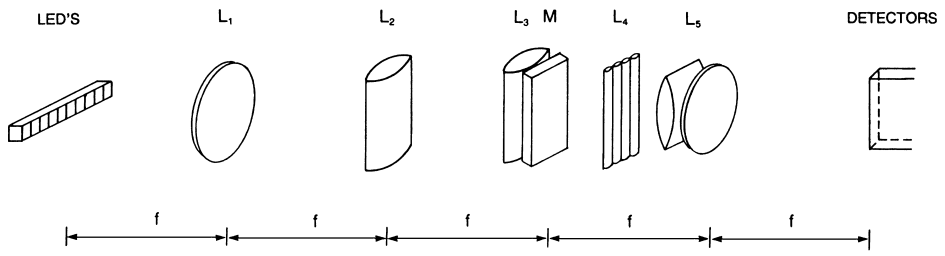


FIGURE 26 Schematic diagram of a matrix–vector multiplier, where f is the focal length of lenses L_1, L_2, and L_5; L_1, spherical lens; L_2 and L_3 are the cylindrical lenses; L_4 is the cylindrical lenslet array; L_5, integrating lens system; and M, mask. [From Goodman, J. W., Dias, A. R., and Woody, L. M. (1978). *Optics Letters* **2**(1), 1–3.]

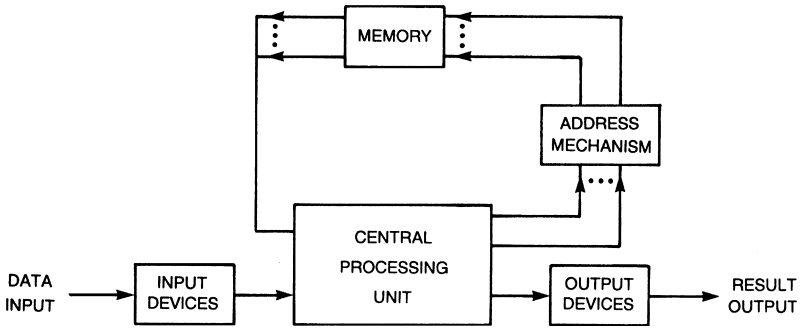


FIGURE 27 Schematic diagram of a von Neumann machine.

it is desirable to perform all the computations optically. Digital computing requires nonlinear processing such as analog-to-digital conversion, logical operations, and bistability. In this section, the general aspects of optical digital computing are described.

A central question is why optical digital computing is needed. The fact is that despite recent advances in electronic digital computers, there exist a number of problem areas such as meteorology, aerodynamics, molecular dynamics, fusion energy, and defense early warning systems that demand computing powers well beyond what is currently available. The major difficulty in increasing the computing power of existing electronic computers is the "von Neumann bottleneck." To illustrate this problem, a schematic diagram of a von Neumann type machine is shown in Fig. 27. Electronic computers generally have this structure. The memory addressing mechanism shown in this figure was proposed by von Neumann in the mid-1940s to decrease the number of connections between the central processing unit and the memory. As an example, consider a memory that has 65,536 binary cells (64-Kbit memory). Direct access to all the cells requires 65,536 wire connections. With binary addressing, the number of connections is reduced to $\log_2 65{,}536 = 16$. In this method, each cell is addressed by a 16-bit binary number. The price that is paid for this reduction is that only one cell can be accessed at a time. This puts a limit on the speed of the processor and it is known as the von Neumann bottleneck. Optical systems do not suffer from such a connection problem. They are inherently two-dimensional, and numerous points can be addressed simultaneously by different beams of light. Therefore, optical digital processors are promising candidates to achieve larger computing capacities.

The idea of building an optical digital computer started in the mid-1960s, when reliable lasers became available. The direction persued at that time was to use the nonlinear properties of materials to perform logic operations. Although the feasibility of this technique was demonstrated, the large amount of power required made the approach impractical. Recently, an increasing demand for parallel processing, advances in electrooptic devices, and the rediscovery of residue arithmetic have made the optical digital computing area active again. A number of different optical systems capable of performing logical operations have been introduced since the 1980s. The basic principles of residue arithmetic are explained in the next subsection. Various approaches for building an optical digital computer are then described, and the problems associated with them are discussed.

B. Residue Arithmetic

Residue arithmetic is a very old subject. The first recorded work in this field is the verse of the Chinese mathematician Sun-Tsu, published around the 1st century *AD*. In this verse, he gives an algorithm for finding a number whose remainders upon division by 3, 5, and 7 are known. A general theory of remainders (now known as the Chinese Remainder Theorem) was established by the German mathematician K. F. Gauss in the 19th century. The application of residue arithmetic in computers, however, is a relatively recent idea. In 1955, the first investigation on this subject was published by A. Svoboda and M. Valach in Czechoslovakia. Their work stimulated more research on residue arithmetic and its applications. Here, a short introduction to residue arithmetic is presented.

Unlike the commonly used decimal and binary number systems, the residue number system (RNS) is an unweighted system. A number system is said to be weighted if any number X can be expressed as $X = \sum_{i=1}^{N} d_i w_i$, where all w_i are constants, all d_i a set of permissible digits, and N is the number of digits used to represent X. Weighted-number systems have favorable features. However, the dependency that exists between the digits puts a limit on the processing speed when an operation is performed by a computer. For example, the addition of two numbers should be done digit by digit, since the addition

of two digits may result in a carry and affect the addition of the next two digits. This is not the case in the RNS.

An RNS is defined by choosing n relatively prime (containing no common factors) numbers $m_1, m_2, \ldots, m_n$ called moduli. Any integer X can then be represented as an n-tuple $(x_1, x_2, \ldots, x_n)$, where $x_i = |X|_{m_i}$ (read X modulo m_i) is the least positive integer remainder that is obtained from the division of X by m_i. For example, if the set of moduli $\{3, 4, 5, 7\}$ is chosen, then the numbers $X = 23$ and $Y = 14$ are represented as $X = (2, 3, 3, 2)$ and $Y = (2, 2, 4, 0)$. In general, given the set of moduli $\{m_1, m_2, \ldots, m_n\}$, a range of $M = \Pi_{i=1}^{n} m_i$ integer numbers can be uniquely represented. This range can be allocated to positive numbers or negative numbers, or it can be partitioned to include both positive and negative numbers.

The important feature of RNS is that basic arithmetic operations can be performed on each digit individually. That is, if $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$ are two numbers of the same system, then $Z = X * Y = (z_1, z_2, \ldots, z_n)$, where $z_i = |x_i * y_i|m_i$, for $i = 1, 2, \ldots, n$, and $*$ represents the addition, subtraction, or multiplication operation. Division is possible, but it is difficult, except for the remainder zero case. As an example, the results of performing addition, subtraction, and multiplication on the two numbers mentioned above are $X + Y = (1, 1, 2, 2)$, $X - Y = (0, 1, 4, 2)$, and $X \cdot Y = (1, 2, 2, 0)$. These are the residue representations of the correct answers, that is, 37, 9, and 322, respectively.

In residue arithmetic, there are a number of basic operations that are difficult to perform. These are division, scaling, sign detection, overflow detection, and relative-magnitude determination. Despite these difficulties, the fact that the calculations associated with different moduli are independent of each other makes RNS suitable for parallel processing. Residue arithmetic is especially well suited for optical computing. One can use the cyclic nature of the phase or polarization of a light beam to implement residue numbers. Also, RNS significantly reduces the storage size required for truth-table look-up processing, described in the next section.

C. Realization of Optical Digital Computers

1. Elements of Optical Digital Computers

The fundamental parts of an optical digital computer are central processing unit, memory, input devices, and output devices. The input devices convert the input data to a form of digital signal suitable for the central processing unit. Among the input devices are analog-to-digital converters, residue converters, and encoders. If the input data are not in the format used by the computer, special devices are used to convert the format. Examples of these devices are serial/electronic-to-parallel/optical converters and serial/optical-to-parallel/optical converters. After the data are processed, the output devices convert the result to a form suitable for detection. Among the output devices are decoders, residue converters, and digital-to-analog converters. Depending on the format of the output signal, the detector may be a single element, or it may be a one-dimensional or a two-dimensional array of elements.

Different types of memories might be required in an optical computer. These include scratch pad memory for storing intermediate results, archival storage unit for storing algorithms, and associative memory for truth-table look-up processing. Among the candidates for optical memories are spatial light modulators, optical disks, holographic memories, and bistable devices. One- or two-dimensional spatial light modulators can be used for temporary storage. Optical disks are suitable for cases that require large storage capacities. Today, optical disks with capacities of 10^{11} bits are commercially available. Volume holographic devices are also suitable for large storage capacities. Storage sizes for $(n/\lambda)^3$ bits/m^3 are theoretically predicted for holographic memories, where λ is the free-space wavelength of the light and n the refractive index of the material used. For a $10 \times 10 \times 2$-mm crystal of $LiNbO_3$ with $n = 2.2$, and with a He–Ne laser with wavelength of 632.8×10^{-9} m, the above storage capacity would be 8.4×10^{12} bits. Due to practical problems, the above limit has not yet been reached. Another advantage of holographic memories is that they can function as associated memories. This means that the whole memory is activated by the input and responds with the associated output. Different beams of light can access the memory simultaneously and perform parallel processing. Another technique for realizing optical memories employs bistable devices. A bistable device has two stable states for an input signal. Hence, it can be used as a binary memory cell to store a logic state of 0 or 1. Electronic bistable devices (flip-flops) have been a major part of electronic digital computers since their invention. Optical bistability was first predicted by A. Szoke and his colleagues at the Massachusetts Institute ofTechnology in 1969. However, due to material problems, they were not successful in showing this effect experimentally. In 1975, H. M. Gibbs and his coworkers at Bell Laboratories demonstrated the effect using sodium vapor as the nonlinear medium. Four years later, the first semiconductor bistable device was built from GaAs by the same group. Today, optical bistability is observed in different materials. Among them, GaAs appears to be the most suitable candidate. With this material, optical bistability has been obtained at room temperature with low powers (<10 mW) and with very high switching times ($\simeq$1 psec switch-on and $\leq$10 nsec switch-off). Further improvements are expected as the parameters are optimized.

The functional heart of an optical digital computer is the central processing unit, which performs various digital arithmetic operations. This unit contains the hardware implementation of the desired operation. It may be made of a single lens to perform Fourier transformation or may include different elements. The techniques for realizing a digital arithmetic operation are described in the following subsection.

2. Methods of Optical Digital Processing

In general, a digital arithmetic operation can be realized in two ways: using logic gates or using truth-table look-up techniques. In the first method, the operation is performed by devices with specific input–output relationships that resemble some arithmetic or logic operations. In the second technique, the result of an operation is read from a memory in which the outputs for all possible input cases are previously stored.

a. Gate-based processing. To realize optical gates, several optical quantities such as amplitude, polarization, and spatial coordinates have been proposed for coding. If amplitude coding is used, various logic operations can be obtained by using transparencies that have controllable transmission. A spatial light modulator that is addressed by an electrical or optical signal can be used as a controlled transparency. Logic operations such as AND or OR can then be implemented by putting these transparencies in cascade or in parallel and applying the inputs as the control signals to these planes. If electrooptic materials are used, the polarization of a beam of light can be changed by applying a voltage. By this technique, optical gates for binary logic or residue arithmetic can be implemented. In the spatial coding technique, different values are represented as the presence of light at a particular set of spatial coordinates. This is particularly suitable for coding residue numbers.

b. Truth-table look-up processing. The use of holography for performing Boolean algebraic operations was first demonstrated by Preston. He recorded Fourier holograms that, upon reconstruction, produced specific phase shifts among the light waves associated with the individual input variables. At the input of this system, the binary states 0 and 1 are modeled as pinhole apertures that introduce phase shifts of 0 and 180° in the laser light, respectively. During processing, the light beams coming out of the input apertures experience specific phase shifts as they are diffracted by the hologram. As a result of the complex amplitude addition of the diffracted waves, the required logic outputs are obtained at the detector plane in the form of bright and dark spots representing 1s and 0s, respectively. Using this technique, Preston performed the logic operation IDENTITY and EXCLUSIVE OR on four pairs of input bits and described how the operation OR could be obtained. Given these, all other logic operations can be constructed.

Preston's idea of using holography for logic operations was expanded by Guest and Gaylord to truth-table look-up processing in their NAND-based processor. Figure 28 shows a simple example of an optical NAND-based processor that implements the multiplication of a pair of two-bit numbers. During the recording process, for each output bit, all the input bit combinations that produce a 1 in that output bit are stored as holograms in a photorefractive crystal such as $LiNbO_3$. During the processing step, the input bits are compared with all the stored reference patterns that correspond to each output bit. If a match is detected, the output bit is considered to be a 1, otherwise it is considered to be a 0. A reference pattern consists of the bits of the two input numbers and an additional bit that is called the reference bit. For example, in the simple case of multiplication of a pair of two-bit numbers, there are four input combinations that produce a 1 in the least significant bit of the output. These are 0101, 0111, 1101, and 1111, which correspond to 01×01, 01×11, 11×01, and 11×11, respectively. Similarly, the number of input combinations that produce a 1 in the second, third, and fourth bits of the output are 6, 3, and 1, respectively. Therefore, a total of $4 + 6 + 3 + 1 = 14$ reference patterns need to be stored, each consisting of five bits (four input bits and one reference bit). Each reference pattern is recorded as a hologram using a reference laser beam at a particular angle of incidence. The required phase shifts for the recording process are obtained by a phase-shifting line composer.

To use the NAND-based processor, the input bits are coded as transparent and opaque regions representing 1s and 0s, respectively. The location of the reference bit is always made transparent. The above coding can be implemented by a binary amplitude line composer. The light passing through the transparent apertures, upon diffraction by the holograms, reconstructs the reference beam at different spatial locations. Depending on the phase of the recorded bits, these diffracted beams are added to or subtracted from each other. For each output bit, if the input pattern matches one of the reference patterns, the diffracted beams cancel each other at the detector array at the location that corresponds to the matched pattern. As a result, a dark spot is produced at that location. The presence of a dark spot is electronically detected, and it is interpreted as a 1 for that particular output bit.

The NAND-based processor described above has several advantages. Input patterns can be entered as two-dimensional arrays, and they can access the whole

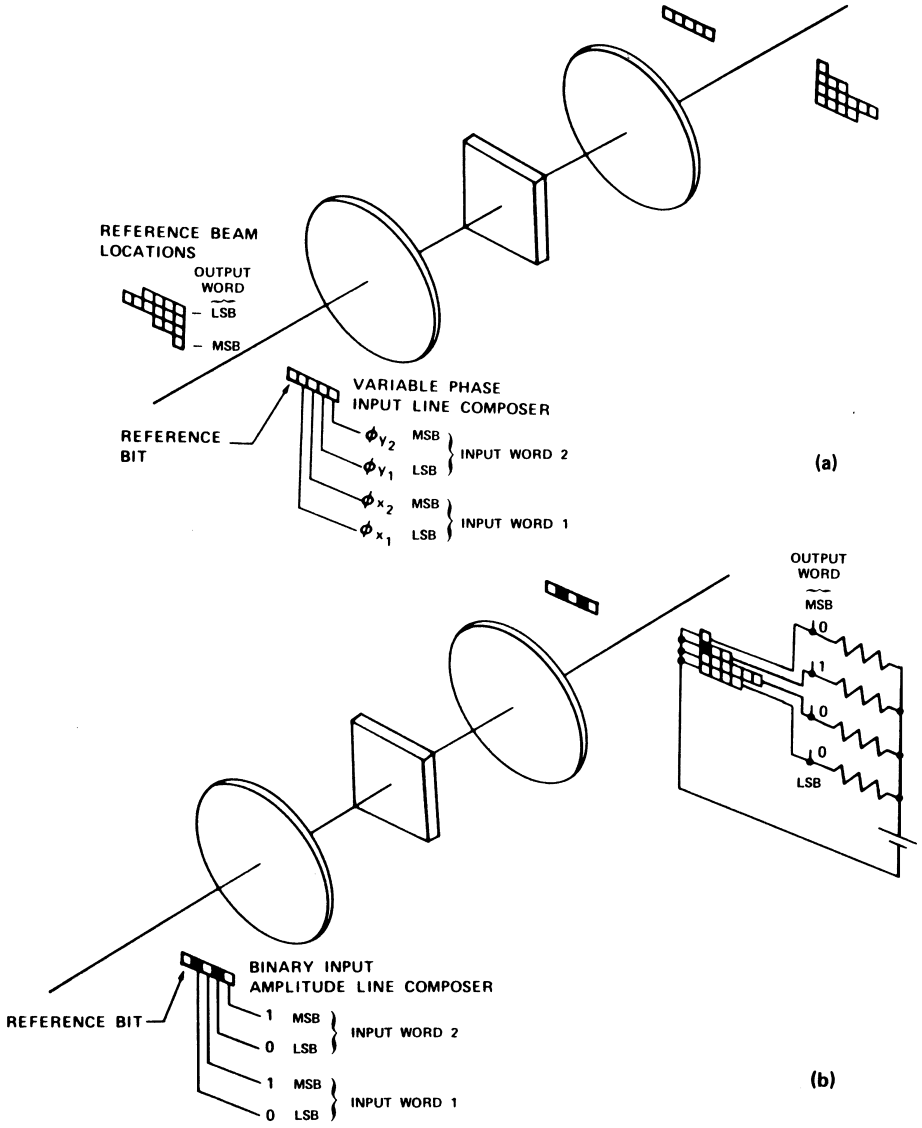


FIGURE 28 NAND-based numerical optical processor: (a) recording the truth-table holograms and (b) example of multiplication with the processor, where LSB is the least significant bit and MSB the most significant bit. [From Guest, C. C., and Gaylord, T. K. (1980). *Applied Optics* **19**(7), 1201–1207.]

memory simultaneously. Therefore, full parallel processing can be achieved. Another advantage of this processor is its flexibility. Any function that can be represented as a truth table could be implemented by this method. Finally, the device is insensitive to the variations in the amplitude and phase of the laser during the processing step. The only limiting problem is the number of reference patterns that can be holographically recorded. The RNS is very useful in decreasing the required number of reference patterns for a particular operation. Further reduction can be obtained if logical minimization techniques and multilevel coding techniques are applied. With these reduction techniques, operations of practical interest, such as 32-bit addition and multiplication, can be implemented by this processor.

3. Current Status of Optical Digital Computing

Research activities in optical computing have significantly increased during the past few years. New techniques have been introduced for optical digital processing. Optical disks are now being used in some electronic computers. Optical interconnections are being investigated for clock distribution and communication between different parts of electronic computers. Also, a new class of optical computers based on neural networks is being studied.

To describe all aspects of optical computing is beyond the scope of this article. Here, we provide a brief description of some of the recent achievements in this field.

a. Optical disks. Today, magnetic hard disks and floppy disks are widely used in electronic computers. A relatively new medium for data storage is optical disks, where the information is recorded and read by a laser beam. The main advantage of optical disks is their high storage capacity. A small 3.5- or 5.5-in. Optical disk is capable of storing 30 to 200 Mbytes of information.

Optical disks are of two types: read-only disks and read–write (erasable) disks. The first type is useful for archival storage and storing data or instructions that do not need to be changed. In the second type, the recorded data can be erased or changed. This type of memory is needed for temporary data storage, such as in digital computing. Some of the materials used for nonerasable disks are tellurium, silver halide, photoresists, and photopolymers. Among the candidate materials for erasable disks, three groups are more promising. These are magneto-optic materials, phase-change materials, and thermoplastic materials.

Optical disks are now used in some models of personal computers, and they are expected to become more common. Also, optical disks have been used for archival storage. Two such systems have been developed and installed by RCA for NASA and Rome Air Development Center in 1985. These are optical disk "jukebox" mass storage systems that provide direct access to any part of a stored data of 10^{13} bits within 6 sec. These systems have a cartridge storage module that contains 125 optical disks, each of 7.8×10^{10} bits storage capacity. This storage size is beyond the capacities that are currently available with other technologies.

b. Optical interconnections. In quantum optics, light is considered to be made of discrete particles known as photons. Unlike electrons, photons do not interact with each other. For example, two beams of light can cross without affecting the paths of each other. This property makes optics very suitable for interconnection.

There are two types of optical interconnections: guided interconnections and free-space interconnections. In the first type, light propagates inside a guiding region such as an optical fiber or an optical waveguide. In the second type, light propagates in the air. Guided interconnections are useful for long-distance interconnections. The application of optical fibers for communication between electronic computers has now been well developed. Free-space interconnections are more suitable for short-distance interconnections, such as the ones needed for communication between different parts of a circuit.

One of the applications of the free-space interconnections, proposed by Goodman in 1984, is the clock distribution. In digital circuits, a clock signal is used to synchronize the operation of different parts of the circuit. With the progress in the technology of integrated circuits, the operating speed and the size of digital circuits continuously increase. As a result, the on-time arrival of the clock signals to all parts of a circuit becomes more and more difficult. An optical signal can be used to overcome this delay problem. One possible architecture is to use a hologram to deflect a beam of light to a number of detectors on the circuit. These detectors convert the optical signals to electronic signals which are then used as the clock signals.

c. Symbolic substitution. One of the new techniques for optical computing is the symbolic substitution method introduced by A. Huang in 1983. In this method, the input data are spatially coded. For example, the binary values 0 and 1 can be coded as transparent and opaque regions shown in Fig. 29(a). To perform an operation on two numbers, the codes of one number are put on top of the codes of the other number. The processor then recognizes specific patterns in the input data and substitutes them with the patterns that correspond to the result. As an illustrative example, the substitution rules for addition of two bits are shown in Fig. 29(b). The addition of two n-bit numbers is performed by repeating these rules on the input pattern n times.

The symbolic substitution technique is quite suitable for optical processing. By using two-dimensional arrays of spatial light modulators, the same operation can be performed on several pairs of input data in the same period that is needed for one pair. The symbolic substitution is capable of performing any operation that can be expressed by

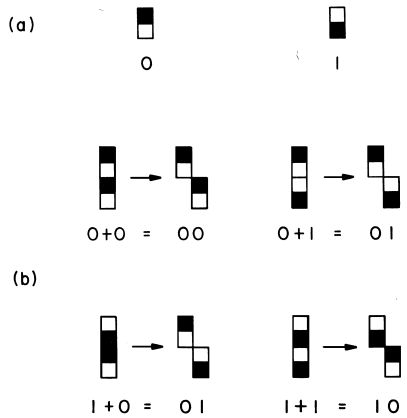


FIGURE 29 (a) Coding procedure for logical values 0 and 1. (b) Substitution rules for addition. [From Brenner, K.-H., and Huang, A. (1985). *Technical Digest of Topical Meeting on Optical Computing*, Optical Society of America, Washington, DC, paper WA4.]

some substitution rules. Also, this technique is not limited to binary arithmetic, and it can use any number system. In particular, the signed-digit number system, introduced by Avizienis in 1961 for parallel arithmetic, can be used to perform addition and subtraction of two numbers in few steps regardless of the size of the numbers.

d. Optical neurocomputers. Despite significant progress in the technology of digital computers, there are certain tasks, such as image recognition, that cannot be efficiently performed, even with the present supercomputers. The poor performance of computers in these areas is not due to the speed of operation or the size of memory, but it is due to the fact that these problems are not algorithmic. The present computers are good at solving problems that have programmable algorithms for their solutions. The human brain, on the other hand, is very good at nonalgorithmic tasks such as image recognition. Although the details of how the brain functions are not known, some general information about the brain has been discovered in neurobiology. For example, it is known that the human brain is made of a large number ($\approx 10^{12}$) of neurons. Each neuron acts as a simple processing unit and is connected to several other neurons. It is also known that the interconnection between the neurons has a major role in the brain activities.

Based on the general information that has been obtained about the human brain, several models have been proposed for neural computers. A general aspect of all these models is a massive interconnection network among the processing units. As described above, optics is capable of global interconnection. Therefore, it is a good candidate for neural computers. The application of optics for implementing neural computers was first proposed by Psaltis and Farhat in 1985. Several other architectures have been introduced since then. The basic architecture of an optical neurocomputer consists of two-dimensional arrays of spatial light modulators that represent the neurons and an optical interconnection network that connects the neurons to each other. The interconnection network can be utilized by a holographic system. Psaltis has investigated the application of thick holograms, where the information is stored in the volume of a recording medium such as a photorefractive crystal. Another approach proposed by Caulfield is to use a two-dimensional array of holograms to realize the interconnection. These and other architectures are currently under investigation.

4. Future of Optical Computing

Although it is hard to predict in detail, it is commonly believed that the role of optics in computing will increase in the future. The short-term impact of optics will probably be to enhance the performance of electronic computers. Fiber-optic interconnections are now being used for communication between computers. In the near future, optical interconnections will also be used for communication at lower levels, such as board-to-board and chip-to-chip. Also, with future progress in the technology of erasable materials, optical disks will become more common in electronic computers.

The long-term impact of optics will be in developing new computers which are not feasible by the existing architectures. The fact that in optics, two-dimensional arrays of data can be processed in parallel makes optics suitable for developing parallel machines and neurocomputers. Many problems should be solved before these machines can become practical. Perhaps the most crucial point is the progress in materials needed for optical devices. Among the devices that need to be improved are spatial light modulators, integrated optical devices, and optical bistable devices.

VII. APPLICATIONS

The applications of optical information processing are so diverse that even brief descriptions of them are not possible here. Some of the areas in which optical processors have been used are antenna pattern analysis, crystal structure analysis, image enhancement, matched filtering, metrology, missile guidance, nondestructive testing, ocean surface analysis, pattern recognition, signal processing, and synthetic aperture radar. Here, we briefly cover two application areas in which optical processing techniques have made significant contributions in their development: synthetic aperture radar and character recognition.

A. Synthetic Aperture Radar

In conventional radar, an electromagnetic pulse is transmitted through the air from a directive antenna. If an airplane is present in the path, the wave is reflected from it and is received by the same antenna. From the time delay between the received pulse and the transmitted pulse, the distance between the antenna and the target can be calculated. Also, from similar measurements at different instants and along different directions, the shape of the target and its velocity can be determined. Theoretically, if the antenna is carried by an airplane, the same method can be used to get information about a target on the ground or to obtain a map of the terrain. However, for high-resolution radar imagery, impractically large antennas are required.

The geometry of an airborne side-looking radar is shown in Fig. 30. It is assumed that the airplane moves with

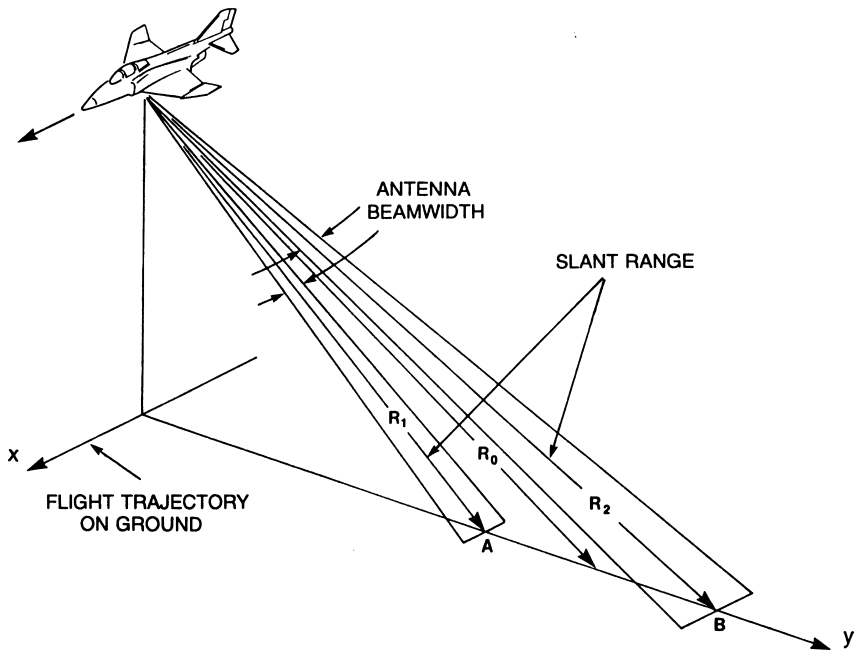


FIGURE 30 Synthetic-aperture radar geometry. [From Preston, K., Jr. (1972). "Coherent Optical Computers," McGraw–Hill, New York.]

a constant velocity along the x axis. Since the slant-range is determined from the delay time between the transmitted and the received signals, the slant-range resolution can be improved by using shorter pulses. Mathematically, the ground-range resolution corresponding to the slant-range resolution can be expressed as follows:

$$\Delta l = c\tau/(2\cos\psi), \tag{74}$$

where c is the velocity of light, τ the pulse duration, and ψ the depression angle of the line of sight to the target with respect to the local horizontal axis. The factor of two in the denominator is due to the fact that the electromagnetic wave travels the same path twice. For a pulse of 1-nsec duration, the maximum ground-range resolution corresponding to $\psi = 0^\circ$ is 15 cm. The along-track or azimuthal resolution depends on the antenna pattern. To get high azimuthal resolution, the half-power angular width of the antenna pattern β should be narrow. This quantity depends on the wavelength of the electromagnetic wave λ and the largest dimension of the aperture of the antenna D. A good approximation is the following equation:

$$\beta \simeq \lambda/D, \tag{75}$$

where λ and D are expressed in the same unit of length and β is in radians. As an illustrative example, if the airplane is 10 km from the target, an azimuthal resolution of 15 cm is obtained with a half-power angular width of $\beta = 0.15/10^4 = 1.5 \times 10^{-5}$ radians. For a wavelength of 3 cm, the required dimension of the aperture antenna is $D = 3 \times 10^{-2}\ \text{m}/1.5 \times 10^{-5} = 2000\ \text{m}$, which is obviously impractical.

The synthetic-aperture technique was developed in the 1950s to improve the azimuthal resolution of an airborne radar. The idea is to synthesize the effect of a large antenna from the data obtained with a small antenna. To get some intuition about this technique, notice that in Fig. 30, a point at range R_0 is illuminated by the pattern of the antenna while the antenna moves a distance of $L = \beta R_0$. The physical antenna, which has an aperture of dimension D, can be considered as an element of a large antenna of length L. If the pulse rate of the radar is such that the antenna does not move more than D during the period between the succeeding pulses, the response of the long antenna can be determined from the gathered data. This is actually a trade-off between time and space. With this technique, an along-track resolution of D can be achieved by processing the gathered data. The process consists of amplitude weighting, phase-shifting, and adding the results. Also, to realize the ultimate along-track resolution for large values of L, it is required to compensate for the curvature of the reflected wavefronts. This focusing operation is range-dependent, and it complicates the process. The amount of information that needs to be stored and the number of computations that are required are so large that real-time analysis could not be achieved even with the most powerful electronic computers. Various methods for storing and analyzing the data were proposed during the early 1950s. Among them, the optical processing

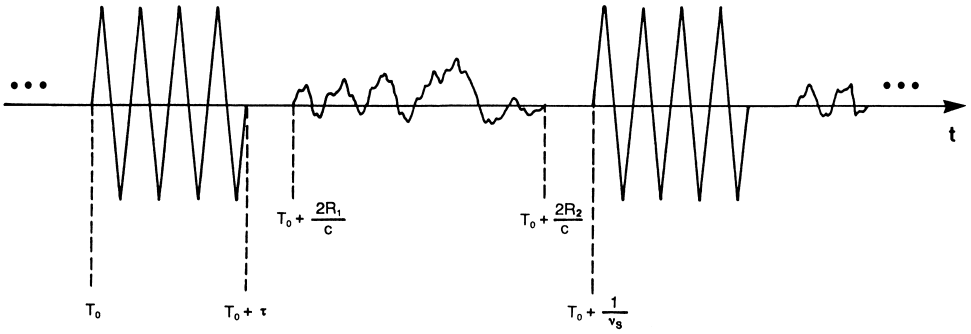


FIGURE 31 Outgoing and incoming side-looking radar pulses, where R_1 and R_2 represent the nearest and farthest slant range probed; τ, pulse duration; and ν_s, pulse rate. [From Brown, W. M., and Porcello, L. J. (1969). *IEEE Spectrum* **6,** 52–62.]

approach was found to be the most suitable technique. Finally, in 1957, Cutrona and his colleagues at the University of Michigan successfully demonstrated a synthetic aperture radar using a coherent optical data processor. Their system is briefly described here.

The radar transmits short pulses of microwave energy of duration τ and rate ν_s, as shown in Fig. 31. The delay relation between the outgoing and the incoming pulses is shown in this figure. The incoming signal is demodulated in the receiver using a reference oscillator similar to the one used for transmission. Since the demodulated signal is bipolar, a dc bias level is added to the signal to make it monopolar. This positive signal is then used to modulate an electron beam created by a cathode-ray tube while the beam is swept along a straight line, as shown in Fig. 32. Each sweep corresponds to one return signal and is recorded along a vertical line on a photographic film. The film moves along the horizontal direction, so that different return signals are recorded at different lines. The variable x_1 is therefore a measure of the along-track position while y_1 corresponds to the slant range. The exposed film is developed such that after process, its amplitude transmittance is a linear function of the amplitude of the demodulated signal. This film is then used as the input to the optical system shown in Fig. 33. In this system, a conical lens is used to insert a range-dependent quadratic phase delay to perform focusing operation. The cylindrical and spherical lens combination takes the Fourier transform of the focused signal along the x direction and images the input along the y direction. The output can be observed through a vertical slit. If a photographic film is placed behind the slit and is moved with the same speed as the data film, a high-resolution two-dimensional picture of the terrain illuminated by the radar is obtained.

B. Character Recognition

One of the applications of optical information processing that clearly demonstrates the parallel processing capability of optical systems is character recognition. This is a special case of a broader class of processes known as pattern recognition. In pattern recognition, the locations at which a particular pattern appears in an input data page are determined. In character recognition, it is desired to locate a particular character or word in an input text page. To perform this search, one technique is to use matched filters. From communication theory, if the impulse response of a system is denoted by $\mathbf{h}(x)$, a filter whose impulse response

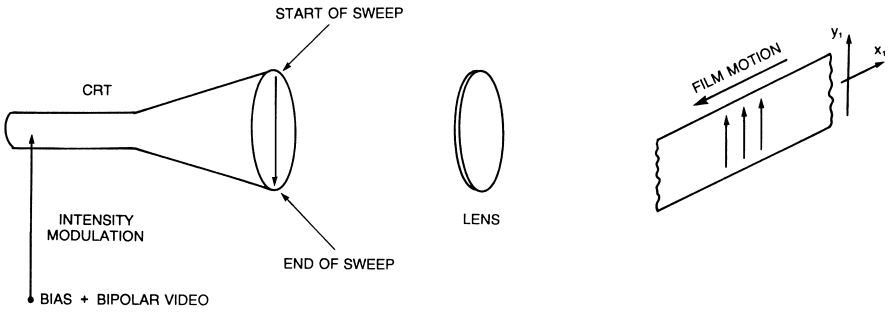


FIGURE 32 Recording process in synthetic aperture radar. [From Brown, W. M., and Porcello, L. J. (1969). *IEEE Spectrum* **6,** 52–62.]

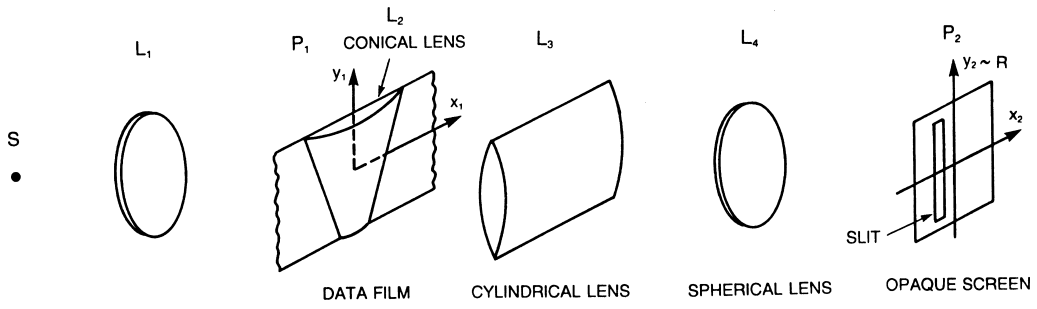


FIGURE 33 Optical processor for synthetic aperture radar analysis. [From Brown, W. M., and Porcello, L. J. (1969). *IEEE Spectrum* **6,** 52–62.]

is $\mathbf{h}^*(-x)$, where * represents the complex conjugate, is known as the corresponding matched filter. Matched filters are used for signal detection. It can be shown that if a signal is buried in a noise that has a uniform spectrum (white noise), the matched filter corresponding to that signal provides the best detection. In character recognition, the signal of interest is a particular character, and it is two-dimensional. The unwanted characters are considered to be noise.

An optical system capable of character recognition can be implemented using a Vander Lugt filter. First, a Fourier hologram of the character or word of interest is recorded using the system shown in Fig. 13a. The transmittance of this hologram has the desired term $\mathbf{H}^*(f_X, f_Y)$, which is the Fourier transform of $\mathbf{h}^*(-x, -y)$. The recorded hologram is used in the Fourier plane of Fig. 12. The input text is placed at the P_1 plane, and the output is detected at the P_3 plane. As described in Section III.B.3, the output of a Vander Lugt filter has four terms. In character recognition, the desired term is the cross-correlation of the input text with the character of interest. In the output plane, at the locations that the desired character appears, the cross-correlation becomes an autocorrelation, and bright spots are produced; at the other parts of the page, the light is diffracted, and these regions are less bright. Example of the impulse response and the output of the processor are shown in Fig. 34. The advantage of optical processing for character recognition is evident from the fact that the search in the input data is performed fully in parallel.

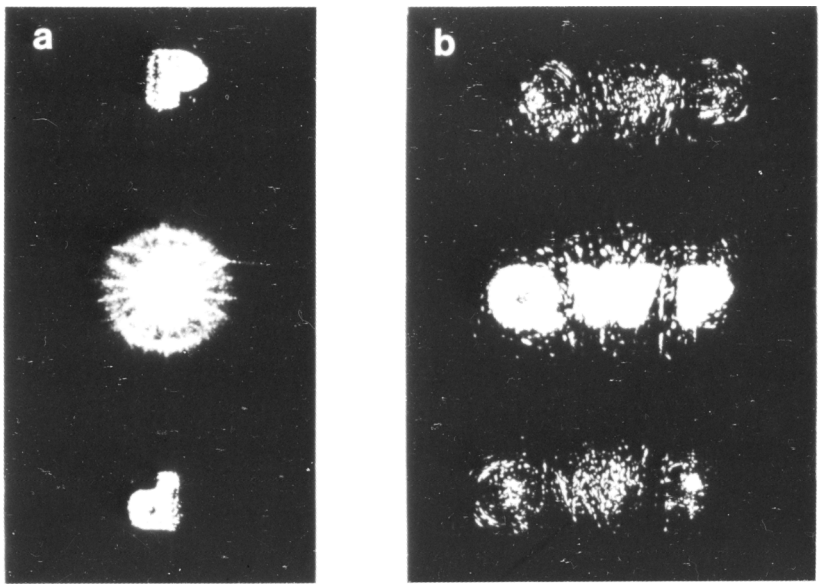


FIGURE 34 Photographs of (a) the impulse response of a Vander Lugt filter and (b) the response of the matched-filter portion of the output to letters Q, W, and P. [From Goodman, J. W. (1968). "Introduction to Fourier Optics," McGraw–Hill, New York.]

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTO-OPTICS • DATA TRANSMISSION MEDIA • HOLOGRAPHY • NONLINEAR OPTICAL PROCESSES • OPTICAL DIFFRACTION • PHOTOGRAPHIC PROCESSES AND MATERIALS POLARIZATION AND POLARIMETRY • RADAR • WAVE PHENOMENA

BIBLIOGRAPHY

Abu-Mostafa, Y. S., and Psaltis, D. (1987). Optical neural computers. *Scientific American* **256,** 88–95.

Arrathoon, R., ed. (1986). "Optical Computing: Digital and Symbolic," Marcel Dekker, New York.

Arsenault, H. H., Szoplik, T., and Macukow, B. (1989). "Optical Processing and Computing," Academic Press, New York.

Bell, T. E. (1986). "Optical computing: A field in flux," *IEEE Spectrum* **23,** 34–57.

Brown, W. M., and Porcello, L. J. (1969). "An introduction to synthetic aperture radar," *IEEE Spectrum* **6,** 52–62.

Caulfield, H. J., and Gheen, G., eds. (1989). "Selected Papers on Optical Computing," SPIE Optical Engineering Press, Bellingham, WA.

Elion, H. A., and Morozov, V. N. (1984). "Optoelectronic Switching Systems in Telecommunications and Computers," Marcel Dekker, New York.

Gibbs, H. M. (1985). "Optical Bistability: Controlling Light with Light," Academic Press, New York.

Goodman, J. W. (1968). "Introduction to Fourier Optics," McGraw–Hill, New York.

Jenkins, F. A., and White, H. E. (1976). "Fundamentals of Optics," McGraw–Hill, New York.

Karim, M. A. (1990). "Electro-optical Devices and Systems," PWS-Kent, Boston.

Lee, S. H., ed. (1981). "Optical Information Processing Fundamentals," Springer-Verlag, New York.

Mirsalehi, M. M., Abushagur, M. A. G., and Caulfield, H. J. (1989). Optical and optoelectronic computing. *In* "Advances in Computers" (M. C. Yovits, ed.), Vol. 28, pp. 153–226, Academic Press, New York.

Preston, K., Jr. (1972). "Coherent Optical Computers," McGraw–Hill, New York.

Stark, H., ed. (1982). "Applications of Optical Fourier Transforms," Academic Press, New York.

Szabo, N. S., and Tanaka, R. I. (1967). "Residue Arithmetic and Its Applications in Computer Technology," McGraw–Hill, New York.



Partially Coherent Processing

Francis T. S. Yu
Pennsylvania State University

I. Introduction
II. Optical Processing under Partially Coherent Regime
III. White-Light Optical Processing
IV. Propagation of Mutual Intensity Function
V. Source Encoding and Image Sampling
VI. Noise Performance
VII. Applications
VIII. Concluding Remarks

GLOSSARY

Achromatic lens Lens without longitudinal chromatic aberration.
Coherent noise Noise generated due to highly coherent light.
Coherent processor Signal processor that utilizes coherent light.
Correlation Mathematical operation involving integration of two functions that are mutual translated with respect to each other.
Diffraction grating Periodic structure transparency to diffract light.
Impulse response Response from a linear system with a very short duration of excitation.
Image sampling An image transparency is on–off sampling by a linear periodic grating.
Interconnection weight matrix Matrix in which the elements represent the interconnection strengths between the processing units.
Interpattern association Association among patterns; the relationship that reflects the common and special features among patterns.
Mutual intensity functions Mutual coherence function defined at zero delay time (i.e., $\tau = 0$).
Narrow spectral band filter Spatial filter with a small finite-spectral bandwidth.
Neural network Network of interconnected processing units that model the behavior of the neurobiological systems.
Partially coherent processor Signal processor that utilizes partially coherent light.
Pseudo-color image False color image.
Source encoding Spatial source sampling.
Spatial coherence Transversal correlation that describes the mutual coherence between any two points on a transversal space (e.g., a plane).
Speech spectrogram Plot of frequency versus time to describe a spectral content of a speech.
Temporal coherence Longitudinal correlation that

describes the mutual coherence between any two points along the longitudinal direction of wave propagation.

White-light processor Optical signal processor that utilizes a white-light source.

PARTIALLY COHERENT PROCESSING is an optical signal processing technique that utilizes the partially coherent properties of light instead of a strict coherent light. The light may be either spatially or temporally partially coherent.

I. INTRODUCTION

Although coherent optical processors can perform a myriad of complicated signal processings, coherent processing systems are usually plagued with coherent artifact noise. These difficulties have prompted us to look at optical processing from a new standpoint: to consider whether it is necessary for all optical processing operations to be carried out by pure coherent sources. We have found that many optical processings can be carried out by partially coherent light or white light. The basic advantages of partially coherent processing are (1) the suppressing of coherent artifact noise; (2) that the partially coherent sources are usually inexpensive; (3) that the constraints on the processing environment are generally very relaxed; (4) that the partially coherent system is relatively easy and economical to operate; and (5) that the partially coherent processor is particularly suitable for color image processing.

II. OPTICAL PROCESSING UNDER PARTIALLY COHERENT REGIME

Optical systems under the partially coherent regime have been studied since 1967. These studies have shown that there are difficulties in applying the linear system concept to the evaluation of the performance at high spatial frequencies. These difficulties are primarily due to the inapplicability of the linear system theory under a partial coherence regime. However, under the strictly coherent regime (i.e., spatially and temporally coherent), the optical system operates in a complex wave field. The output complex light field can be described by a complex amplitude convolution integral, such as

$$g(\alpha, \beta) = \iint_{-\infty}^{\infty} f(x, y)h(\alpha - x, \beta - y)\,dx\,dy \qquad (1)$$

which yields a spatially invariant property of the complex amplitude transformation, where $f(x, y)$ is a coherent complex wave field at the input end of an optical system, and $h(x, y)$ describes the spatial impulse response of the optical system. Equation (1) can be written in the Fourier transform form

$$G(p, q) = F(p, q)H(p, q) \qquad (2)$$

where $G(p, q)$, $H(p, q)$, and $F(p, q)$ are the Fourier transforms of $g(x, y)$, $h(x, y)$, and $f(x, y)$, respectively. Thus Eq. (2) describes the optical system operation in the spatial frequency domain, where $H(p, q)$ can be referred to as the *coherent* or *complex amplitude transfer function* of the optical system.

On the other hand, if the optical system is under a strictly incoherent regime, the system operates in intensity, such that the output irradiance is described by the intensity convolution integral

$$I(\alpha, \beta) = \iint_{-\infty}^{\infty} f_i(x, y)h_i(\alpha - x, \beta - y)\,dx\,dy \qquad (3)$$

which yields a spatially invariant form of the intensity operation, where $f_i(x, y) = |f(x, y)|^2$ is the input irradiance and $h_i(x, y) = |h(x, y)|^2$ is the intensity spatial impulse response of the optical system. Similarly, Eq. (3) can be written in Fourier transform form, that is,

$$I(p, q) = F_i(p, q)H_i(p, q) \qquad (4)$$

where $I(p, q)$, $F_i(p, q)$, and $H_i(p, q)$ are the Fourier transforms of $I(x, y)$, $f_i(x, y)$, and $h_i(x, y)$, respectively, and $H_i(p, q)$ can be referred to as the *incoherent* or *intensity transfer function* of the optical system.

We shall now define the following normalized quantities:

$$\tilde{I}_i(p, q) = \frac{\displaystyle\iint_{-\infty}^{\infty} I(\alpha, \beta)\exp[-i(p\alpha + q\beta)]\,d\alpha\,d\beta}{\displaystyle\iint_{-\infty}^{\infty} I(\alpha, \beta)\,d\alpha\,d\beta} \qquad (5)$$

$$\tilde{F}_i(p, q) = \frac{\displaystyle\iint_{-\infty}^{\infty} f_i(x, y)\exp[-i(px + qy)]\,dx\,dy}{\displaystyle\iint_{-\infty}^{\infty} f_i(x, y)\,dx\,dy} \qquad (6)$$

$$\tilde{H}_i(p, q) = \frac{\displaystyle\iint_{-\infty}^{\infty} h_i(x, y)\exp[-i(px + qy)]\,dx\,dy}{\displaystyle\iint_{-\infty}^{\infty} h_i(x, y)\,dx\,dy} \qquad (7)$$

These normalized quantities would give a set of convenient mathematical forms and also provide a concept of image contrast interpretation. Since the quality of a visual image depends to a large extent on the contrast (or the relative irradiance) of the image, a normalized intensity function would certainly enhance the information bearing capacity. With the application of the Fourier convolution theorem to Eq. (3), the following relationship can be written:

$$\tilde{I}_i(p,q) = \tilde{F}_i(p,q)\tilde{H}_i(p,q) \tag{8}$$

where $\tilde{H}_i(p,q)$ is commonly referred to as the *optical transfer function* (OTF) of the optical system, and the modulus $|\tilde{H}(p,q)|$ is known as the *modulation transfer function* (MTF) of the optical system.

We further note that Eq. (7) can also be written as

$$\tilde{H}_i(p,q) = \frac{\displaystyle\int\int_{-\infty}^{\infty} H(p',q')H^*(p'-p,q'-q)\,dp'\,dq'}{\displaystyle\int\int_{-\infty}^{\infty} |H(p',q')|^2\,dp'\,dq'} \tag{9}$$

By changing the variables $p'' = p' - p/2$, $q'' = q' - q/2$, Eq. (9) results in the following symmetrical form:

$$\tilde{H}_i(p,q) = \frac{\displaystyle\int\int_{-\infty}^{\infty} H(p''+p/2,q''+q/2) \times H^*(p''-p/2,q''-q/2)\,dp''\,dq''}{\displaystyle\int\int_{-\infty}^{\infty} |H(p'',q'')|^2\,dp''\,dq''} \tag{10}$$

We note that the definition of OTF is valid for any linearly spatial invariant optical system, regardless of whether the system is with or without aberrations. Furthermore, Eq. (10) serves as the primary link between the strictly coherent and strictly incoherent systems. There are, however, threefold limitations associated with the strictly coherent processing. First, coherent processing requires the dynamic range of a spatial light modulator to be about 100,000:1 for the input wave intensity, or a photographic density range of 4.0. Such a dynamic range is often quite difficult to achieve in practice. Second, coherent processing systems are susceptible to coherent artifact noise, which frequently degrades the image quality. Third, in coherent processing, the signal being processed is carried out by a complex wave field. However, what is actually measured at the output plane of the optical processor is the output wave intensity. The loss of the output phase distribution may seriously limit the applicability of the optical system in some applications.

The drawbacks of coherent processing, can be overcome by reducing either the temporal coherence or the spatial coherence, and there are several techniques available that use those approaches. In the following, we briefly mention four frequently used concepts. The first method is based on a spatially partially coherent source, and the other three methods utilize a broad spectral band light source (e.g., a white-light source) to perform complex wave field operations.

A. Spatially Partially Coherent Processing

Techniques of utilizing a spatially partially coherent light to perform complex data processing have been proposed. These techniques share a basic concept: the optical system is characterized by the use of a point spread function (PSF), which is not constrained to the class of nonnegative real functions used in conventional incoherent processing. The output intensity distribution can be adjusted by changing the PSF, that is,

$$I(x',y') = \int\int O(x,y)h(x',y';x,y)\,dx\,dy \tag{11}$$

where $h(x,y;x',y')$ is the PSF, and $I(x',y')$ and $O(x,y)$ are the image and object intensity distributions, respectively. If the pupil function of an optical system is $P(\alpha,\beta)$, the PSF is equal to the square of the Fourier spectrum of the pupil function. Therefore, the output intensity distribution can be adjusted by selecting an adequate pupil function. A typical processing system, as proposed by Rhodes, is shown in Fig. 1. This optical processing system, which is characterized by an extended pupil region, has two input pupil functions, $P_1(\alpha,\beta)$ and $P_2(\alpha,\beta)$. With the optical path lengths of the two arms of the system being equal, the overall system pupil function is given by the sum of $P_1(\alpha,\beta)$ and $P_2(\alpha,\beta)$. However, if the path length in one arm is changed slightly, for example, by moving mirror

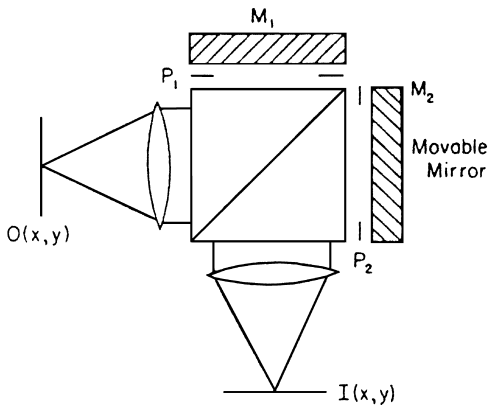


FIGURE 1 A two-pupil incoherent processing system.

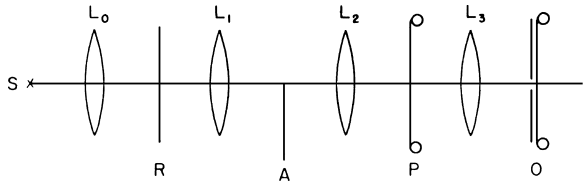


FIGURE 2 An achromatic optical processing system.

M_2 a small distance, a phase factor is introduced in one of the component pupil functions, with the result

$$P(\alpha, \beta) = P_1(\alpha, \beta) + P_2(\alpha, \beta) \exp[i\phi(\alpha, \beta)] \qquad (12)$$

It is evident that if the pupil transparencies $P_1(\alpha, \beta)$ and $P_2(\alpha, \beta)$ are recorded holographically, arbitrary PSFs can then be synthesized using the 0–180° phase-switching operation. Thus we see that this optical system is capable of performing complex data processing with spatially partially coherent light.

B. Achromatic Optical Processing

An achromatic optical signal processing technique with broad-band source (e.g., white light) is shown in Fig. 2. Lens L_0 collimates the white-light point source S illuminating hologram R. The diffraction wavefront is spatially filtered by the aperture A, which removes the zero-order term and the lower sideband. The upper sideband is imaged onto the signal plane P; lens L_3 Fourier-transforms the resulting wavefront; and observation in the transform plane is confined to the optical axis by a slit in the output plane O. The convolution of the demodulated input signal with the desired reference function is recorded by synchronously translating the signal and output films. Such a system, termed an achromatic system, thus has the flexibility of a coherent optical signal processor, along with the potential for the noise immunity of an incoherent system.

C. Band-Limited Partially Coherent Processing

A technique of a matched filter for operation with band-limited illumination is shown in Fig. 3. The matched filter consists of a frequency-plane holographic filter, an achromatic-fringe interferometer, a color-compensating grating, and an achromatic doublet. The matched filter, a Fourier hologram, is made by recording the object spectrum and a collimated reference beam with an exposure wavelength λ_0. The reference-beam angle is at θ_0 with respect to the object beam axis. In the correlation operation, the object is illuminated using a broad-spectral source. The various spectral components are dispersed in angle due to the grating-like structure of the matched filter. A lateral dispersed correlation signal is generated in plane c, which is the conventional correlation plane. These spectral components are recombined by imaging the matched filter into plane d, where a compensation grating produces a color dispersion that is equal but opposite to that introduced at plane b. The color-corrected correlation signal is observed in the output plane. However, the compensated signal is still wavelength-dependent, since the illumination wavelength changes the scale of the object spectrum at plane b. Insertion of a slit in plane c provides a convenient way to bandlimit the correlation signal to $\Delta\lambda$. Bandlimiting improves the signal-to-noise ratio (SNR) of the compensated correlation output. Broadband illumination can be used for automatic scale search and object size determination. Since spatially incoherent light as well as coherent light can be used to perform matched filtering, an extended white-light source is used together with a slit at plane c to provide a bandlimited partially coherent processing.

D. Achromatic Partially Coherent Processing

An achromatic partially coherent optical processing technique with a white-light source was introduced by Yu in 1978. Figure 4 illustrates the processing system, where all the transform lenses are assumed achromatic. A high-diffraction-efficiency phase grating with an angular spatial

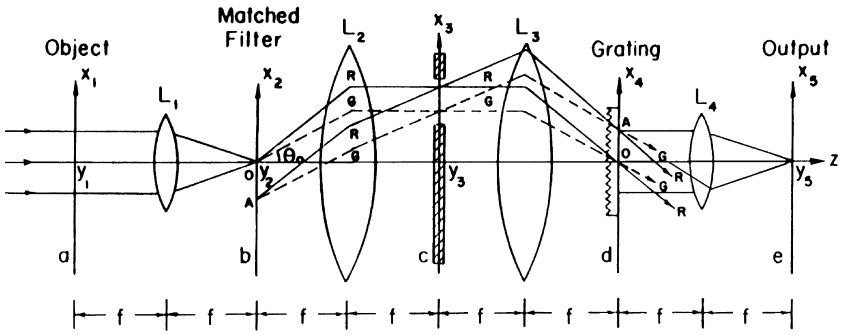


FIGURE 3 Band-limited partially coherent processing system.

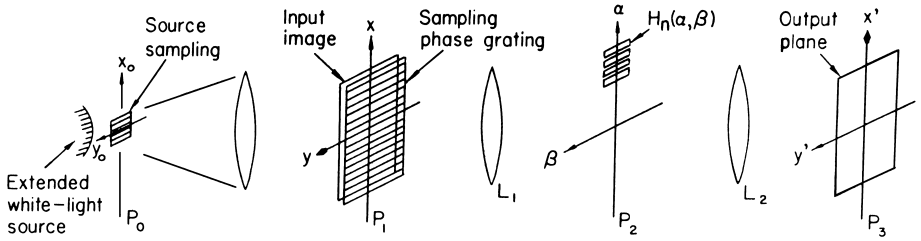


FIGURE 4 An achromatic partially coherent optical processing system.

frequency P_0 is used at the input plane P_1 to disperse the input object spectrum into rainbow color in the Fourier plane P_2. Thus it is permitting a stripwise design of a complex spatial filter for each narrow spectral band in the Fourier plane. The achromatic output image irradiance can therefore be observed at the output plane P_3. The advantages of this technique are that each channel (e.g., each spectral band filter) behaves as a partially coherent channel, while the overall output noise performance is primarily due to the incoherent addition of each channel, which behaves as if under incoherent illumination. Thus, this partially coherent system has the capability for coherent noise suppression.

III. WHITE-LIGHT OPTICAL PROCESSING

We shall now describe a partially coherent processor that can be carried out by a white-light source, as shown in Fig. 4. The partially coherent processing system is similar to a coherent processing system, except that it uses an extended white-light source, a source encoding mask, a signal sampling grating, multispectral band filters, and achromatic transform lenses. For example, if we place an input object transparency $s(x, y)$ in contact with a sampling phase grating, the complex wave field, for every wavelength λ, at the Fourier plane P_2 would be (assuming a white-light point source)

$$E(p, q; \lambda) = \iint s(x, y) \exp(ip_0 x) \times \exp[-i(px + qy)]dxdy = S(p - p_0, q) \tag{13}$$

where the integral is over the spatial domain of the input plane P_1; (p, q) denotes the angular spatial frequency coordinate system; p_0 is the angular spatial frequency of the sampling phase grating; and $S(p, q)$ is the Fourier spectrum of $s(x, y)$. If we write Eq. (13) in the form of a spatial coordinate system (α, β), we have

$$E(\alpha, \beta; \lambda) = S\left(\alpha - \frac{\lambda f}{2\pi} p_0, \beta\right) \tag{14}$$

where $p = (2\pi/\lambda f)\alpha$, $q = (2\pi/\lambda f)\beta$, and f is the focal length of the achromatic transform lens. Thus we see that the Fourier spectra would disperse into rainbow colors along the α axis, and each Fourier spectrum for a given wavelength λ is centered at $\alpha = (\lambda f/2\pi)p_0$.

In complex spatial filtering, we assume that a set of narrow-spectral-band complex spatial filters is available. In practice, all the input objects are spatial-frequency-limited. The spatial bandwidth of each spectral band filter $H(p_n, q_n)$ is therefore

$$H(p_n, q_n) = \begin{cases} H(p_n, q_n) & \alpha_1 < \alpha < \alpha_2 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $p_n = (2\pi/\lambda_n f)\alpha$, $q_n = (2\pi/\lambda_n f)\beta$, λ_n is the main wavelength of the filter, $\alpha_1 = (\lambda_n f/2\pi)(p_0 + \Delta p)$ and $\alpha_2 = (\lambda_n f/2\pi)(p_0 - \Delta p)$ are the upper and lower spatial limits of $H(p_n, q_n)$, and Δp is the spatial bandwidth of the input image $s(x, y)$.

Since the limiting wavelengths of each $H(p_n, q_n)$ are

$$\lambda_l = \lambda_n \frac{p_0 + \Delta p}{p_0 - \Delta p} \quad \text{and} \quad \lambda_h = \lambda_n \frac{p_0 - \Delta p}{p_0 + \Delta p} \tag{16}$$

its spectral bandwidth can be approximated by

$$\Delta\lambda_n = \lambda_n \frac{4p_0(\Delta p)}{p^2 - (\Delta p)^2} \approx \frac{4(\Delta p)}{p_0}\lambda_n \tag{17}$$

If we place this set of spectral band filters side by side and position them properly over the smeared Fourier spectra, the intensity distribution of the output light field can be shown as

$$I(x, y) \approx \sum_{n=1}^{N} \Delta\lambda_n |s(x, y; \lambda_n) * h(x, y; \lambda_n)|^2 \tag{18}$$

where $h(x, y; \lambda_n)$ is the spatial impulse response of $H(p_n, q_n)$ and $*$ denotes the convolution operation. Thus, the proposed partially coherent processor is capable of processing the signal in a complex wave field. Since the output intensity is the sum of the mutually incoherent narrow-band spectral irradiances, the annoying coherent artifact noise can be suppressed.

IV. PROPAGATION OF MUTUAL INTENSITY FUNCTION

In investigating the behavior of a partially coherent processor, we would establish a transformational relationship of the mutual intensity function, which would determine the coherence requirement of a partially coherent processor. We shall use Wolf's theory (from 1955) to develop a transformational formula for the white-light optical system.

One of the most remarkable and useful properties of a converging lens is its inherent ability to perform a two-dimensional Fourier transformation of a complex wave field. Under partially coherent illumination, a thin lens would take a four-dimensional Fourier transformation of the mutual intensity function. Evidently, when a monochromatic plane wave passes through a thin lens, phase-delay transformation takes place, such as

$$T(\xi, \eta) = \exp(ik\eta\Delta_0)\exp[-ik(\xi^2+\eta^2)/2f] \qquad (19)$$

where η is the refractive index of the thin lens, f is the focal length, Δ_0 is the maximum thickness of the lens, $k = 2\pi/\lambda$, λ is the wavelength, and (ξ, η) is the spatial coordinate system of the thin lens.

Let us now consider an object transparency, inserted at a distance d_0 in front of the lens and illuminated by a spatially partially coherent light. If the mutual intensity function at the object plane is $J_1(x_1, y_1; x_2, y_2)$, as depicted in Fig. 5, then the mutual intensity function at the output (α, β) plane can be shown as

$$\begin{aligned} &J_4(\alpha_1, \beta_1; \alpha_2, \beta_2) \\ &= C\exp\left(i\frac{k}{2d_1}\left\{1-\frac{1}{\alpha_1[(1/d_1)+(1/d_0)-(1/f)]}\right\}\right. \\ &\quad\left.\times\left[(\alpha_1^2+\beta_1^2)-(\alpha_2^2+\beta_2^2)\right]\right) \\ &\quad\times\int_{-\infty}^{\infty}\iiint J_1(x_1, y_1; x_2, y_2) \\ &\quad\times\exp\left(i\frac{k}{2d_0}\left\{1-\frac{1}{d_0[(1/d_1)+(1/d_0)-(1/f)]}\right\}\right. \\ &\quad\left.\times\left[(x_1^2+y_1^2)-(x_2^2+y_2^2)\right]\right) \\ &\quad\times\exp\left\{-ik\frac{(\xi_1x_1+\eta_1y_1)-(\xi_2x_2+\eta_2y_2)}{d_0d_1[(1/d_1)+(1/d_0)-(1/f)]}\right\} \\ &\quad\times dx_1\,dy_1\,dx_2\,dy_2 \end{aligned} \qquad (20)$$

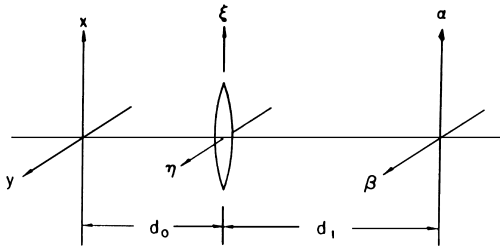


FIGURE 5 A partially coherent optical system.

where C is an appropriate constant. This equation shows that the output mutual intensity function can be obtained by a four-dimensional integral equation.

Up to this point we have disregarded the finite extent of the lens aperture. Such an approximation is an accurate one, if the distance d_0 is sufficiently small to place the input transparency deep within the region of Fresnel diffraction with respect to the lens aperture. This condition is satisfied in the vast majority of problems of interest, particularly for optical processing.

There are, however, two special cases worth mentioning:

1. For $d_1 = f$, the output plane is located at the back focal plane. Equation (20) reduces to

$$\begin{aligned} &J_4(\alpha_1, \beta_1; \alpha_2, \beta_2) \\ &= C\exp\left\{i\frac{k}{2f}\left(1-\frac{d_0}{f}\right)\left[(\alpha_1^2+\beta_1^2)-(\alpha_2^2+\beta_2^2)\right]\right. \\ &\quad\left.\times\int_{-\infty}^{\infty}\iiint J_1(x_1, y_1; x_2, y_2)\exp\left\{-i\frac{k}{2f}[(\alpha_1x_1+\beta_1y_1)\right.\right. \\ &\quad\left.\left.-(\alpha_2x_2+\beta_2y_2)]\right\}dx_1\,dy_1\,dx_2\,dy_2 \end{aligned} \qquad (21)$$

Thus, except for a quadratic phase factor, the output mutual intensity function is essentially the Fourier transform of the input mutual intensity function.

2. For $d_0 = d_1 = f$, the input and output planes are located at the front and back focal length of the lens. The quadratic phase factor vanishes and Eq. (21) reduces to

$$\begin{aligned} &J_4(\alpha_1, \beta_1; \alpha_2, \beta_2) \\ &= C\int_{-\infty}^{\infty}\iiint J_1(x_1, y_1; x_2, y_2)\exp\left\{-i\frac{k}{2f}[(\alpha_1x_1+\beta_1y_1)\right. \\ &\quad\left.-(\alpha_2x_2+\beta_2y_2)]\right\}dx_1\,dy_1\,dx_2\,dy_2 \end{aligned} \qquad (22)$$

which is exactly a four-dimensional Fourier transformation, between the input and output mutual intensity functions.

In partially coherent processing, an extended incoherent source is usually used at the front focal plane of a

collimating lens to illuminate the input object transparency. Thus the mutual intensity function at the source plane (x_0, y_0) can be written as

$$J(x_0', y_0'; x_0'', y_0'') = \begin{cases} \gamma(x_0, y_0) & \text{for} \quad x_0' = x_0'' = x_0, \\ & \qquad y_0' = y_0'' = y_0 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where $\gamma(x_0, y_0)$ is the intensity distribution of the light source, and (x_0, y_0) is the spatial coordinate system of the source plane. If we assume that the input object transparency is located at the back focal plane of the collimator, then the mutual intensity function at the input plane reduces to

$$J_1(x_1 - x_2; y_1 - y_2) = \int_{\infty}^{\infty}\int \gamma(x_0, y_0) \times \exp\left\{-i\frac{k}{2f}[(x_1 - x_2)x_0 + (y_1 - y_2)y_0]\right\} dx_0\, dy_0 \tag{24}$$

which is essentially the Van Cittert–Zernike theorem. It is evident that the mutual intensity function at the input plane is a spatially invariant function.

A. General Formulation

With reference to the partially coherent processor of Fig. 4, the mutual intensity function at the input plane P_1 due to the source irradiance $\gamma(x_0, y_0; \lambda)$ is

$$J(x_1, y_1; x_2, y_2; \lambda) = \iint \gamma(x_0, y_0; \lambda) \times \exp\left\{-i\frac{2\pi}{\lambda f}[(x_1 - x_2)x_0 + (y_1 - y_2)y_0]\right\} dx_0\, dy_0 \tag{25}$$

where the integration is over the source plane P_0. The mutual intensity function immediately behind the sampling phase grating can be written as

$$J'(x_1 y_1; x_2, y_2; \lambda) = J(x_1, y_1; x_2, y_2; \lambda) s(x_1, y_1) s^*(x_2, y_2) \times \exp(i2\pi\nu_0 x_1) \exp(-i2\pi\nu_0 x_2) \tag{26}$$

where the superscript * denotes the complex conjugate, and ν_0 is the spatial frequency of the sampling grating. Similarly, the mutual intensity function at the Fourier plane P_2 can be written as

$$J(\alpha_1, \beta_1; \alpha_2, \beta_2; \lambda) = \iint \gamma(x_0, y_0; \lambda) S(x_0 + \alpha_1 - \lambda f \nu_0, y_0 + \beta_1) \times S^*(x_0 + \alpha_2 + \lambda f \nu_0, y_0 + \beta_2)\, dx_0\, dy_0 \tag{27}$$

where the integration is over the source plane P_0, and $S(\alpha, \beta)$ is the Fourier spectrum of $s(x, y)$. If we assume that a set of narrow-spectral-band spatial filters $H_n(\alpha, \beta)$ are inserted at the Fourier plane, the output mutual intensity function would be

$$J(x_1', y_1'; x_2', y_2'; \lambda) = \iiiint J'(x_1, y_1; x_2, y_2; \lambda) \times \exp\left[-i\frac{2\pi}{\lambda f}(x_1'\alpha_1 + y_1'\beta_1 - x_2'\alpha_2 - y_2'\beta_2)\right] d\alpha_1\, d\beta_1\, d\alpha_2\, d\beta_2 \tag{28}$$

Thus the corresponding output intensity distribution is

$$I_n(x', y'; \lambda) = \iint \gamma(x_0, y_0; \lambda) \times \left| \iint S(x_0 + \alpha - \lambda f \nu_0, y_0 + \beta) H_n(\alpha, \beta) \times \exp\left[-i\frac{2\pi}{\lambda f}(\alpha x' + \beta y')\right] d\alpha\, d\beta \right|^2 dx_0\, dy_0 \quad \text{for} \quad \lambda_{ln} \leqq \lambda \leqq \lambda_{hn} \tag{29}$$

where λ_{ln} and λ_{hn} are the lower and upper wave length limits of $H_n(\alpha, \beta)$. Let us denote $S(\lambda)$ and $C(\lambda)$ as the relative spectral intensity of the light source and the relative spectral response of the output detector. The output irradiance resulting from each narrow spectral band filter would then be

$$I_n(x', y') = \int_{\lambda_n - \Delta\lambda_n/2}^{\lambda_n + \Delta\lambda_n/2} \iint \gamma(x_0, y_0; \lambda) S(\lambda) C(\lambda) \times \left| \iint S(x_0 + \alpha - \lambda f \nu_0, y_0 + \beta) H_n(\alpha, \beta) \times \exp\left[-i\frac{2\pi}{\lambda f}(x'\alpha + y'\beta)\right] d\alpha\, d\beta \right|^2 dx_0\, dy_0\, d\lambda \quad \text{for} \quad n = 1, 2, \ldots, N \tag{30}$$

where λ_n and $\Delta\lambda_n$ are the center wavelength and the bandwidth of the nth narrow-spectral-band filter $H_n(\alpha, \beta)$, respectively. It is therefore apparent that the overall output irradiance would be the incoherent addition of $I_n(x', y')$, that is,

$$I(x', y') = \sum_{n=1}^{n} I_n(x', y') \tag{31}$$

where N is the total number of the spectral-band filters. Furthermore, if the image processing is a one-dimensional operation, then a fan-shaped spatial filter can be utilized; thus overall output irradiance is reduced to the form

$$I(x', y') = \iiint \gamma(x_0, y_0; \lambda) S(\lambda) C(\lambda) \times \left| \iint S(x_0 + \alpha - \lambda f \nu_0, y_0 + \beta) H(\alpha, \beta) \times \exp\left[-i \frac{2\pi}{\lambda f}(x'\alpha + y'\beta)\right] d\alpha\, d\beta \right|^2 dx_0\, dy_0\, d\lambda \tag{32}$$

It is evident that the coherence requirement for partially coherent processing can be determined with either Eq. (31) or Eq. (32).

V. SOURCE ENCODING AND IMAGE SAMPLING

We shall now discuss a linear transform relationship between the spatial coherence (i.e., mutual intensity function) and the source encoding. Since the spatial coherence depends on the image processing operation, a more relaxed coherence requirement may be used for specific image-processing operations. The object of source encoding is to alleviate the stringent coherence requirement so that an extended source can be used. In other words, the source encoding is capable of generating an appropriate spatial coherence for a specific optical signal processing, such that the available light power from the source may be efficiently utilized.

A. Source Encoding

We begin our discussion with Young's experiment under an extended source illumination, as shown in Fig. 6. First, we assume that a narrow slit is placed in the source plane P_0 behind an extended monochromatic source. To maintain a high degree of coherence between the slits Q_1 and Q_2 at plane P_2, the source size should be very narrow. If the separation between Q_1 and Q_2 is large, then a narrower slit size S_1 is required. Thus the slit width should be

$$w \leqq \frac{\lambda R}{2h_0} \tag{33}$$

where R is the distance between planes P_0 and P_1, and $2h_0$ is the separation between Q_1 and Q_2. Let us now consider two narrow slits S_1 and S_2 located in source plane P_0. We

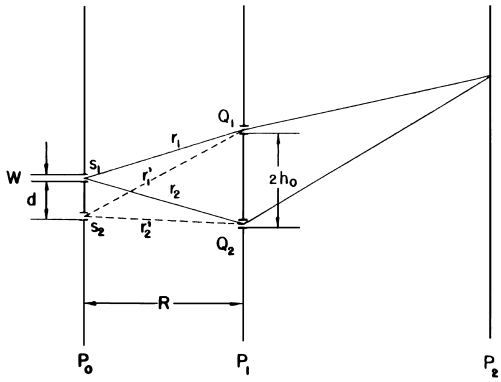


FIGURE 6 Concept of source encoding.

assume that the separation between S_1 and S_2 satisfies the path-length relation

$$r_1' - r_2' = (r_1 - r_2) + m\lambda \tag{34}$$

where the r's are the respective distances from S_1 and S_2 to Q_1 and Q_2, m is an arbitrary integer, and λ is the wavelength of the extended source. Then the interference fringes due to each of the two source slits S_1 and S_2 should be in phase, and a brighter fringe pattern can be seen at plane P_2. To further increase the intensity of the fringes, one would simply increase the number of slits in appropriate locations in plane P_0 such that the separation between slits satisfies the fringe condition of Eq. (34). If the separation R is large, that is, if $R \gg d$ and $R \gg 2h_0$, then the spacing d becomes

$$d = m \frac{\lambda R}{2h_0} \tag{35}$$

Thus by properly encoding an extended source, it is possible to maintain a high degree of coherence between Q_1 and Q_2 and, at the same time, to increase the intensity of the fringes.

To encode an extended source, we would first search for a coherence function for a specific image processing operation. With reference to the partially coherent processor of Fig. 4, the mutual intensity function at the input plane P_1 can be written as

$$J(\mathbf{x}_1, \mathbf{x}_1') = \iint \gamma(\mathbf{x}_0) K(\mathbf{x}_0, \mathbf{x}_1) K^*(\mathbf{x}_0, \mathbf{x}_1')\, d\mathbf{x}_0 \tag{36}$$

where the integration is over the source plane P_0, superscript * denotes the complex conjugation, $\gamma(x_0)$ is the intensity distribution of the encoding mask, and $K(\mathbf{x}_0, \mathbf{x}_1)$ is the transmittance function between the source plane P_0 and the input plane P_1, which can be written

$$K(\mathbf{x}_0, \mathbf{x}_1) \approx \exp\left(i 2\pi \frac{\mathbf{x}_0 \mathbf{x}_1}{\lambda f}\right) \tag{37}$$

By substituting $K(\mathbf{x}_0, \mathbf{x}_1)$ into Eq. (36), we have

$$J(\mathbf{x}_1 - \mathbf{x}_1') = \iint \gamma(\mathbf{x}_0) \exp\left[i2\pi \frac{\mathbf{x}_0}{\lambda f}(\mathbf{x}_1 - \mathbf{x}_1')\right] d\mathbf{x}_0 \quad (38)$$

From the above equation, we see that the spatial coherence and source encoding intensity form a Fourier transform pair, that is,

$$\gamma(x_0) = f[J(\mathbf{x}_1 - \mathbf{x}_1')] \quad (39)$$

and

$$J(\mathbf{x}_1 - \mathbf{x}_1') = f^{-1}[\delta(\mathbf{x}_0)] \quad (40)$$

where f denotes the Fourier transformation operation. It is evident that the relationship of Eqs. (39) and (40) is the well-known Van Cittert–Zernike theorem. In other words, if a required spatial coherence is given, then a source encoding transmittance can be obtained through the Fourier transformation. In practice, however, the source encoding transmittance should be a positive real quantity that satisfies the physically realizability condition:

$$0 \leqq \gamma(\mathbf{x}_0) \leqq 1 \quad (41)$$

B. Image Sampling

There is, however, a temporal coherence requirement for partially coherent processing. If we restrict the Fourier spectra, due to wavelength spread, within a small fraction of the fringe spacing d of a narrow spectral band filter $H_n(\alpha, \beta)$, then we have

$$\frac{P_m f \Delta\lambda_n}{2\pi} \ll d \quad (42)$$

where $1/d$ is the highest spatial frequency of the filter, P_m is the angular spatial frequency limit of the input image transparency, f is the focal length of the achromatic transform lens, and $\Delta\lambda_n$ is the spectral bandwidth of $H_n(\alpha, \beta)$. The temporal coherence requirement of the spatial filter is, therefore,

$$\frac{\Delta\lambda_n}{\lambda_n} \ll \frac{\pi}{h_0 P_m} \quad (43)$$

where λ_n is the central wavelength of the nth narrow-spectral-band filter, and $2h_0 = \lambda_n f/d$ is the size of the input object transparency.

VI. NOISE PERFORMANCE

White-light optical processors are known to perform better under noisy conditions than do their coherent counterparts. We shall now show some effects of the noise performance under white-light illumination.

An experimental setup for the study of noise performance of a partially coherent processor is shown in Fig. 7. We shall first investigate the noise performance under the spatially coherent illumination, that is, the effects due to source size (ΔS). For convenience, we assume that the source irradiance is uniform over a square aperture at source plane P_0, which can be written as

$$\gamma(x_0, y_0) = \text{rect}\left(\frac{x_0}{a}\right) \text{rect}\left(\frac{y_0}{a}\right) \quad (44)$$

where

$$\text{rect}\left(\frac{x_0}{a}\right) \equiv \begin{cases} 1, & |x_0| \leq \dfrac{a}{2} \\ 0, & |x_0| > \dfrac{a}{2} \end{cases}$$

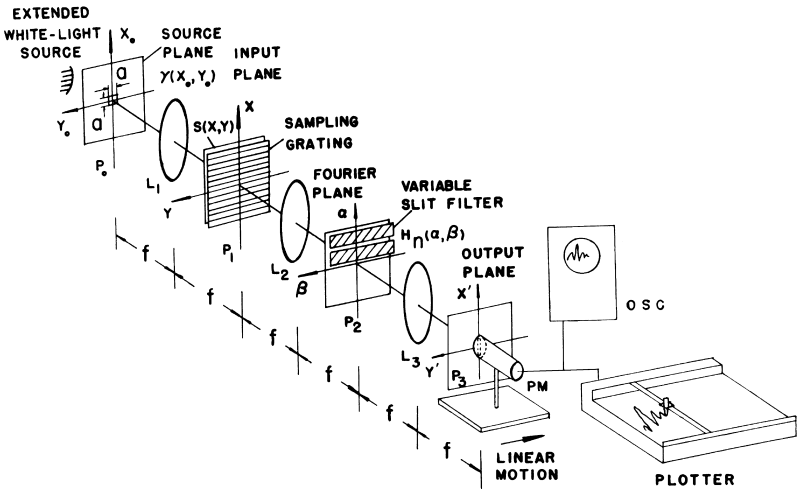


FIGURE 7 Noise measurement of a partially coherent processor: $\gamma(x_0, y_0)$, source intensity distribution; P_0, source plane; P_1, input plane; P_2, input plane; P_3, output plane; $s(x, y)$, object transparency; $H_n(\alpha, \beta)$, slit filters; PM, photometer; OSC, oscilloscope.

For simplicity, we assume that the input signal is a one-dimensional object independent of the y axis. The Fourier spectrum would also be one-dimensional in the β direction, but smeared into rainbow colors along the α direction. Let the width of the nth narrow-spectral-band filter be $\Delta\alpha_n$. If the filter is placed in the smeared Fourier spectra, the spectral bandwidth of the filter can be written as

$$\Delta\lambda_n = \frac{\Delta\alpha_n}{\nu_0 f} \tag{45}$$

The total number of filter channels can therefore be determined as

$$N \equiv \frac{\Delta\lambda}{\Delta\lambda_n} \approx \frac{\Delta\lambda\nu_0 f}{\Delta\alpha_n} \tag{46}$$

where $\Delta\lambda$ is the spectral bandwidth of the white-light source. Thus, we see that the degree of temporal coherence at the Fourier plane increases as the spatial frequency of the sampling grating ν_0 increases.

For the noise performance in the temporally coherent regime, we would use a variable slit representing a broad spectral filter in the Fourier plane. The output noise fluctuation can be traced out with a linearly scanning photometer, as illustrated in Fig. 7. It is apparent that the output noise fluctuation due to the spectral bandwidth of the slit filter and due to the source size can be separately determined. We shall use the following definition of output signal-to-noise ratio (SNR) to evaluate the noise performance of the partially coherent processor:

$$\mathrm{SNR}_n(y') \equiv E[I_n(y')]/\sigma_n(y') \tag{47}$$

where $I_n(y')$ is the output irradiance due to nth channel, E[ ] denotes the ensemble average, and $\sigma_n^2(y')$ is the variance of the output noise fluctuation, that is,

$$\sigma_n^2(y') \equiv E\left[I_n^2(y')\right] - \{E[I_n(y')]\}^2 \tag{48}$$

Evidently, the output intensity fluctuation $I_n(y')$ can be traced out by a linearly scanning photometer. The dc component of the output traces is obviously the output signal irradiance (i.e., $E[I_n]$), and the mean square fluctuation of the traces is the variance of the output noise (i.e., σ_n^2). Thus, we see that the effect of the output SNR due to spectral bandwidth (i.e., temporal coherence) and source size (i.e., spatial coherence) can be readily obtained.

We shall first demonstrate the noise performance due to perturbation at the input plane. For the amplitude noise at the input plane, the experiments have shown that there is no apparent improvement in noise performance under the partially coherent illumination. The result is quite consistent with the 1978 prediction by Chavel and Lowenthal. However, for the phase noise at the input plane, the noise performance of the system is largely improved with a partially coherent illumination. We shall first utilize a weak phase model as an input noise. We consider the situation of an input object transparency superimposed with a thin phase noise at the input plane. The effect on the noise performance of the optical system can then be obtained by varying the source size and the spectral bandwidth of the slit filter. Figure 8 shows a set of output images with photometer traces illustrating the output noise due to the spectral bandwidth of the slit filter. From these pictures, we see that the SNR improves as the spectral bandwidth of the slit filter increases.

A quantitative measurement of the noise performance due to phase noise at the input plane is plotted in Figs. 9 and 10. From these figures, we see that the output SNR increases monotonically as the spectral bandwidth $\Delta\lambda_n$ of the slit filter increases, and is linearly increasing as the source size enlarges. Thus, the noise performance for a partially coherent processor improves as the degree of coherence (i.e., temporal and spatial coherence) relaxes. In other words, to improve the output SNR of the partially coherent processor, one can relax either the spatial coherence, the temporal coherence, or both.

Let us now demonstrate the effect of the noise performance due to strong phase noise. A conventional shower glass was used as an input phase noise. Figure 11 shows a set of results obtained under various spectral bandwidth illuminations. Figure 11a shows an output result obtained under entire broad-band white-light illumination. Although this image is somewhat aberrated due to the thick phase perturbation, the image is relatively immune to random noise fluctuation. Comparing the results obtained from Figs. 11a to 11d, we see that the output SNR decreases rather rapidly as the spectral bandwidth of the slit filter decreases. Figure 11d shows the result obtained with a HeNe laser. Aside from the poor noise performance, we have noted that the output image is severely corrupted by coherent artifact noise.

We shall now demonstrate the noise performance due to noise at the Fourier plane. The effects of amplitude noise at the Fourier plane are plotted in Fig. 12. In contrast with the amplitude noise at the input plane, we see that the output SNR increases monotonically as the spectral bandwidth of the slit filter increases. The output SNR also improves as the source size enlarges. Thus, for amplitude noise at the Fourier plane, the noise performance of a partially coherent processor improves as the degree of temporal and spatial coherence decreases.

Figure 13 shows the noise performance due to phase noise at the Fourier plane. From this figure, once again we see that the output SNR is a monotonically increasing function of the spectral bandwidth $\Delta\lambda_n$. The SNR also increases as the source size increases. However, as compared with the case of phase noise at the input plane of Fig. 10, the improvement of the noise performance is somewhat less effective.

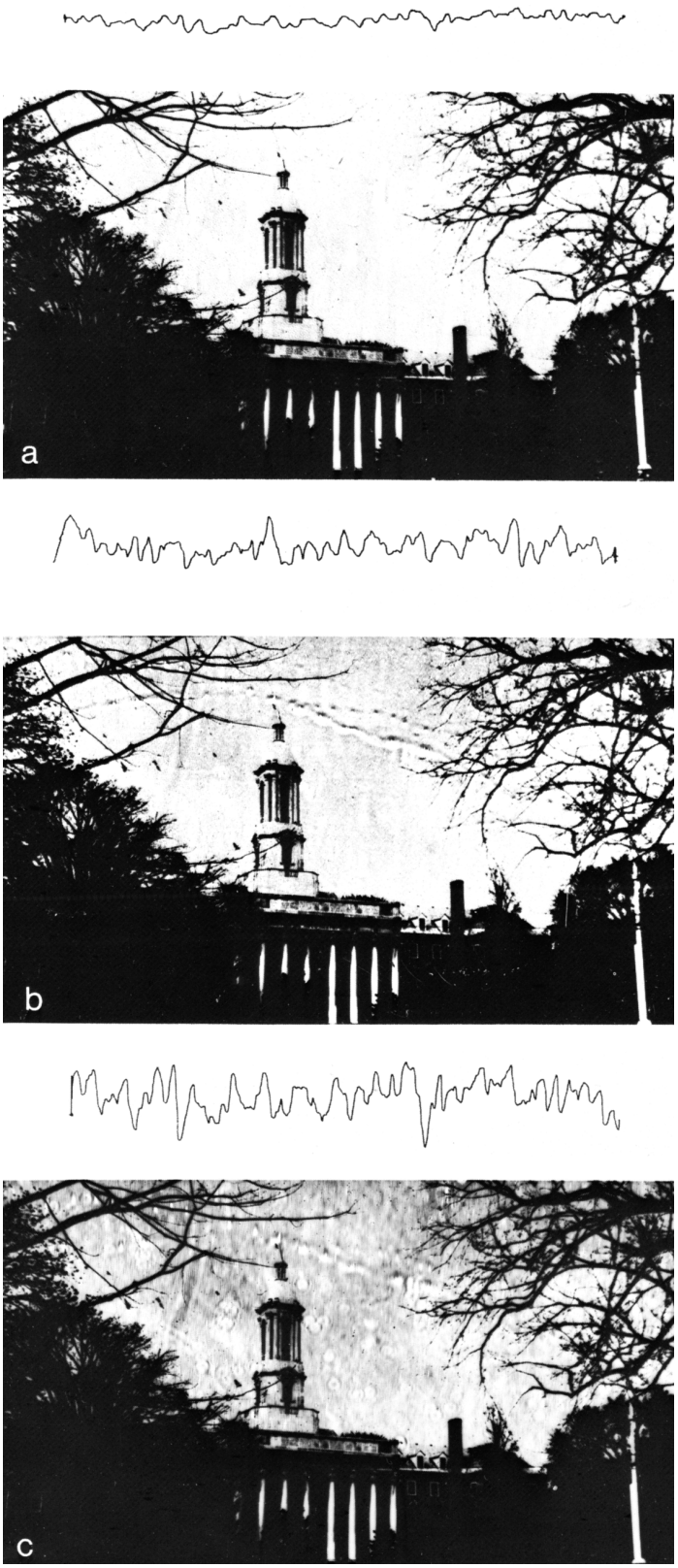


FIGURE 8 Effect on output image (with a section of photometer traces) due to phase noise at input plane for different spectral bandwidths. (a) For $\lambda_n = 1500$ Å. (b) For $\lambda_n = 1000$ Å. (c) For $\lambda_n = 500$ Å.

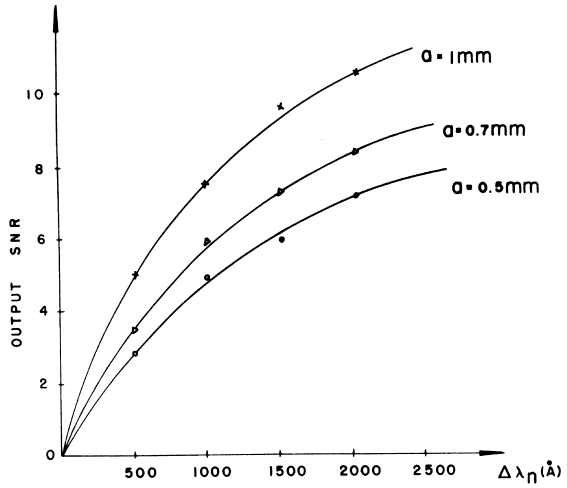


FIGURE 9 Output signal-to-noise ratio for amplitude noise at input plane for various source signs.

We shall now provide the result of noise performance due to thin phase noise along the optical axis (i.e., Z axis) of the optical system. Figure 14 shows the variation of output SNR due to phase noise inserted in various planes of the optical system. From this figure, we see that the output SNR improves drastically for phase noise inserted at the input and at the output plane under temporally and spatially partially coherent illumination. The noise performance is somewhat less effective for the phase noise at the Fourier plane, even under partially coherent illumination. Nonetheless, the phase noise at the Fourier plane can, in principle, be totally eliminated under very-broad-band illumination, if each of the noise channels is uncorrelated. We have also noted that the output SNR is somewhat lower for higher spatial coherent illumination. In other words, the output SNR can also be improved with extended source illumination.

The noise performance due to amplitude noise along the Z axis is plotted in Fig. 15. From this figure, we notice that the output SNR improves as the noise perturbation is moved away from the input and the output plane, and that the optimum SNR occurs at the Fourier plane. Again, we see that the output SNR is somewhat higher for a larger source size (i.e., lower degree of spatial coherence). It is apparent that if the amplitude noise is placed either at the input or at the output plane, the noise performance cannot be improved, under either partially coherent or incoherent illumination.

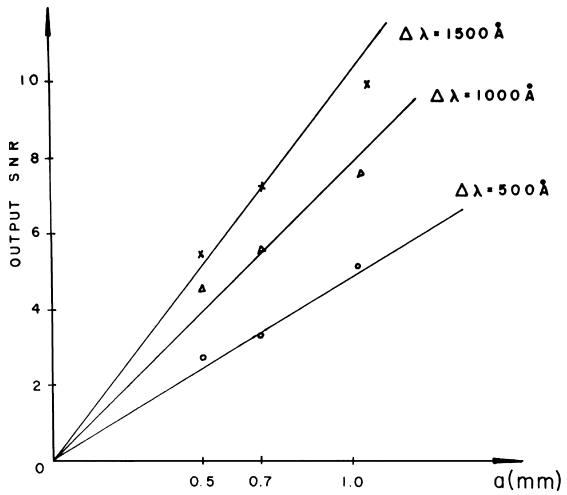


FIGURE 10 Output signal-to-noise ratio for phase noise at input plane for various spectral bandwidths.

VII. APPLICATIONS

It would occupy innumerable pages to describe all the applications of partially coherent processing. In this section, we restrict ourselves to a few applications that are of interest to our readers. As we have noted in the previous discussion, the partially coherent processor is very suitable for color-image processing. Therefore, in the following examples we confine ourselves to color-image processing.

A. Restoration of Blurred Images

One of the interesting applications of partially coherent processing is the restoration of blurred color photographic images. Since a linearly smeared image is a one-dimensional restoration problem and the deblurring operation is a point-by-point filtering process, a *fan-shape* type deblurring filter is used to compensate for the scale variation of the smeared Fourier spectra due to wavelength.

Let us now describe a color-image deblurring technique utilizing a fan-shape deblurring filter. We assume a blurred image due to linear motion described by

$$\hat{s}(x, y) = s(x, y) * \text{rect}\left(\frac{y}{W}\right), \tag{49}$$

where $\hat{s}(x, y)$ and $s(x, y)$ denote the blurred and unblurred images, respectively, and

$$\text{rect}\left(\frac{y}{W}\right) \equiv \begin{cases} 1 & y \le \dfrac{W}{2} \\ 0 & \text{otherwise} \end{cases}$$

where W is the smeared length.

If we insert this blurred image transparency into the input plane P_1 of the partially coherent processor of Fig. 4, the complex light distribution for every wavelength λ at the back focal length of the achromatic transform lens L_2 would be

$$E(\alpha, \beta; \lambda) = S\left[\alpha - \left(\frac{\lambda f}{2\pi}\right) p_0, \beta\right] \text{sinc}\left(\frac{\pi W}{\lambda f}\beta\right) \tag{50}$$

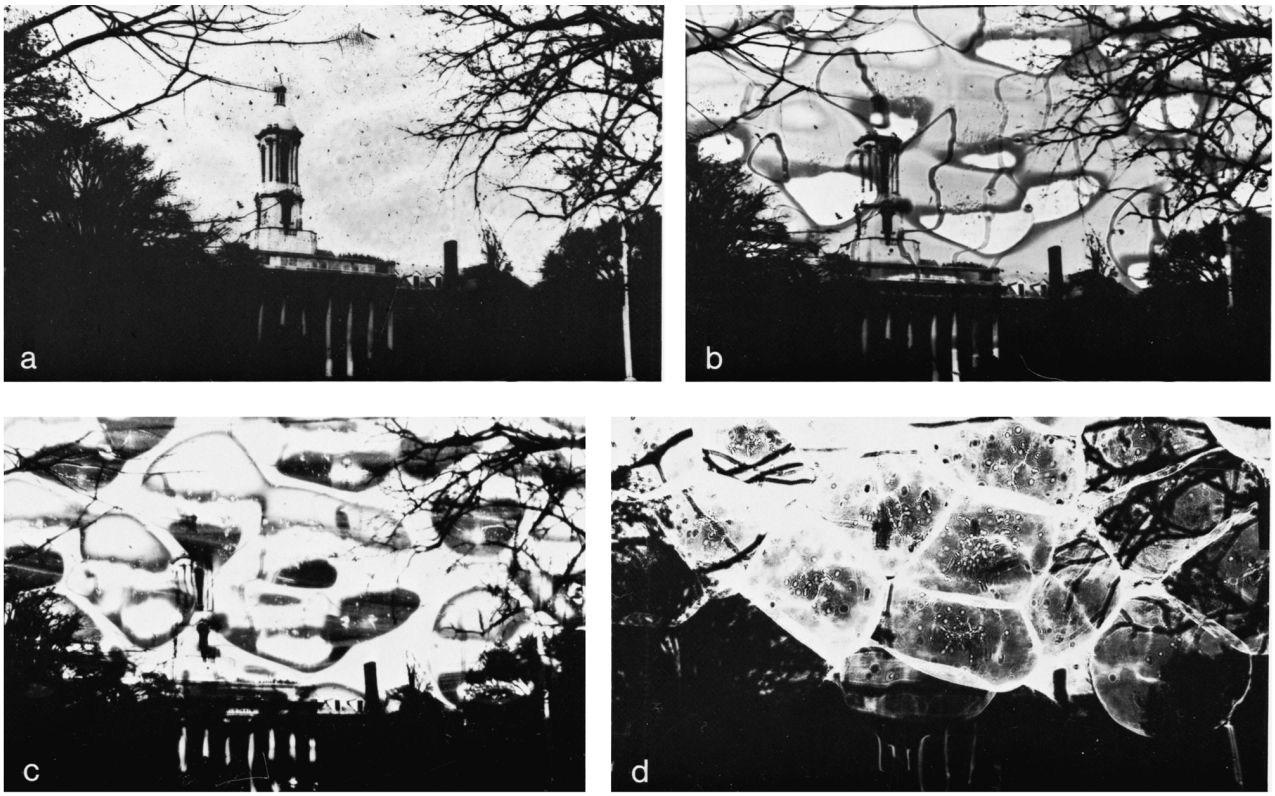


FIGURE 11 Effect on output image due to strong phase perturbation at input plane. (a) For $\Delta\lambda_n = 3000$ Å. (b) For $\Delta\lambda_n = 1500$ Å. (c) For $\Delta\lambda_n = 500$ Å. (d) Obtained with an HeNe laser.

where p_0 is the angular spatial frequency of the phase grating and $S[\cdot]$ represents the smeared Fourier spectrum of $s(x, y)$. Equation (50) shows that the image spectrum is smeared into a rainbow color spectrum.

We use a fan-shaped deblurring filter as described in the following equation:

$$H(\alpha, \beta; \lambda) = \delta\left(\alpha - \frac{\lambda f}{2\pi} p_0, \beta\right)\left\{\int \left[\text{rect}\left(\frac{y}{W}\right) \times \exp\left(-i\frac{2\pi}{\lambda f}\beta y\right)\right] dy\right\}^{-1}$$
$$= \delta\left(\alpha - \frac{\lambda f}{2\pi} p_0, \beta\right)\left[\text{sinc}\left(\frac{\pi W}{\lambda f}\beta\right)\right]^{-1} \tag{51}$$

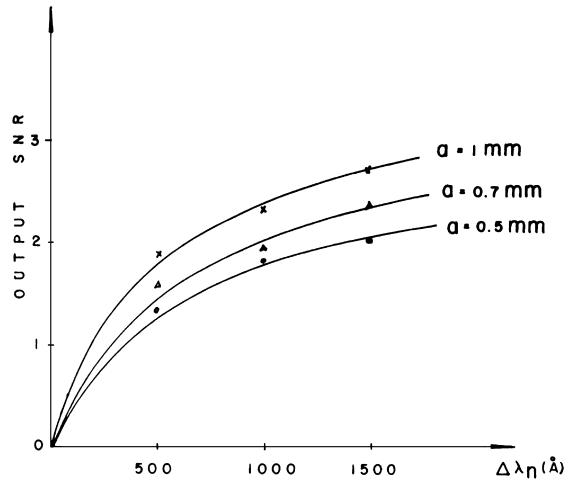


FIGURE 12 Output signal-to-noise ratio for amplitude noise at Fourier plane for various source sizes.

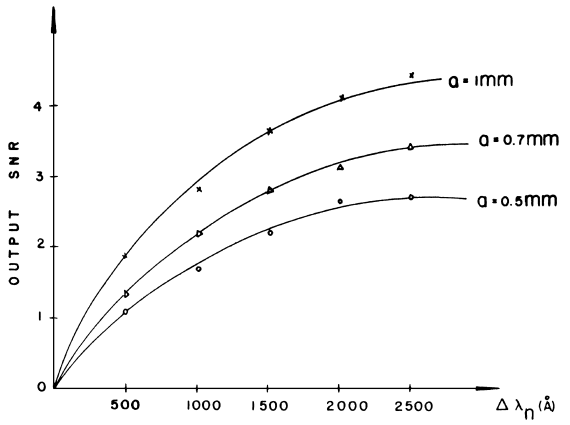


FIGURE 13 Output signal-to-noise ratio for phase noise at Fourier plane for various source sizes.

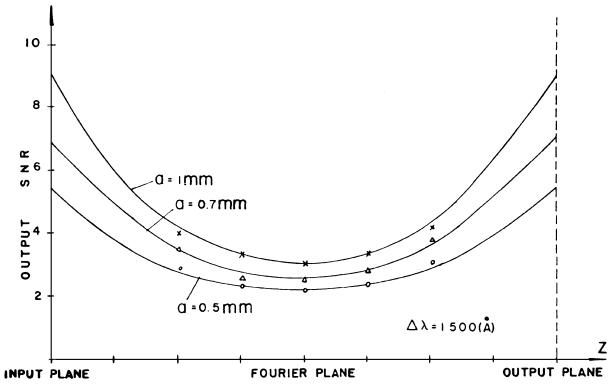


FIGURE 14 Variation of output signal-to-noise ratio due to thin phase noise as a function of the Z direction for various source sizes.

If this deblurring filter is inserted in the spatial frequency plane, the complex light distribution for every λ at the output image plane would be

$$g(x, y; \lambda) = \mathcal{F}^{-1}\left[S\left(\alpha - \frac{\lambda f}{2\pi} p_0, \beta\right) H(\alpha, \beta; \lambda)\right] = s(x, y) \exp(i p_0 x) \tag{52}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. The resultant output image intensity is therefore

$$I(x, y) = \int_{\Delta\lambda} |g(x, y; \lambda)|^2 \, d\lambda \approx \Delta\lambda |s(x, y)|^2 \tag{53}$$

which is proportional to the spectral bandwidth $\Delta\lambda$ of the white-light source. Thus, we see that the partially coherent processor is capable of deblurring color images.

Figure 16a shows a black-and-white picture of a deblurred color image obtained with the white-light processing technique. Compared with the original blurred image of Fig. 16b, we see that the "stop" sign, the posts, the cars, and the trees are far more distinguishable in the deblurred one. The color reproduction of the deblurred image is rather faithful, and the coherent artifact noise is virtually nonexistent.

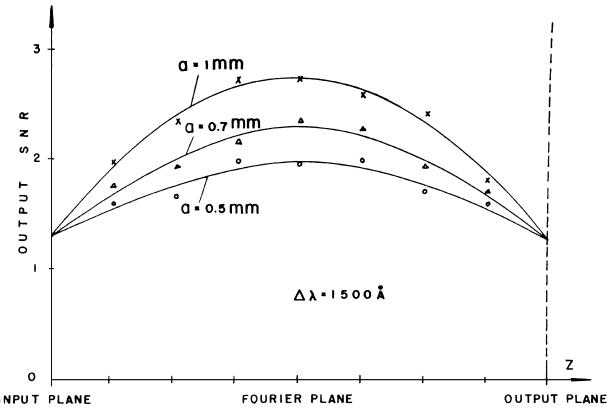


FIGURE 15 Variation of output signal-to-noise ratio due to thin amplitude noise as a function of the Z direction for various source sizes.

B. Image Subtraction

Another interesting application of white-light optical processing is color-image subtraction. Two color-image transparencies are inserted in the input plane of the white-light optical processor of Fig. 4. At the spatial frequency plane, the complex light distribution for each wavelength λ of the light source may be described as

$$E(\alpha, \beta; \lambda) = S_1\left(\alpha - \frac{\lambda f}{2\pi} p_0, \beta\right) \exp\left(-i \frac{2\pi}{\lambda f} h_0 \beta\right) + S_2\left(\alpha - \frac{\lambda f}{2\pi} p_0, \beta\right) \exp\left(i \frac{2\pi}{\lambda f} h_0 \beta\right) \tag{54}$$

where $S_1(\alpha, \beta)$ and $S_2(\alpha, \beta)$ are the Fourier spectra of the input color images $s_1(x, y)$ and $s_2(x, y)$, $2h_0$ is main separation of the two color images, and p_0 is the angular spatial frequency of the sampling phase grating. Again we see that the two input image spectra disperse into rainbow colors along the α axis of the spatial frequency plane.

For image subtraction, we should insert a sinusoidal grating in the spatial frequency plane. Since the scales of the Fourier spectra vary with respect to the wavelength of the light source, we must utilize a fan-shaped grating to compensate for the scale variation. Let us assume the transmittance of the fan-shaped grating is

$$H(\alpha, \beta; \lambda) = \left[1 + \sin\left(\frac{2\pi}{\lambda f} h_0 \beta\right)\right] \quad \text{for all } \alpha \tag{55}$$

Then the output image irradiance is

$$I(x, y) \approx \Delta\lambda \big[|s_1(x, y - h_0)|^2 + |s_2(x, y + h_0)|^2 + \tfrac{1}{2}|s_1(x, y) - s_2(x, y)|^2 + |s_1(x, y - 2h_0)|^2 + |s_2(x, y + 2h_0)|^2\big] \tag{56}$$

where $\Delta\lambda$ is the spectral bandwidth of the white-light source. Thus, the substracted color image can be seen at the optical axis of the output plane. In practice, it is difficult to obtain a true white-light point source. However, this short coming can be overcome with the source encoding technique described in Section V.A.

To ensure a physically realizable source encoding function, we let a spatial coherence function with an appropriate point-pair coherence requirement be



FIGURE 16 Restoration of blurred images. (a) A black-and-white deblurred color picture due to linear motion. (b) A black-and-white blurred color image.

$$\Gamma(|y-y'|)=\frac{\sin[(N\pi/h_0)|y-y'|]}{N\sin[(\pi/h_0)|y-y'|]}\operatorname{sinc}\left(\frac{\pi w}{h_0 d}|y-y'|\right) \tag{57}$$

where N is a positive integer $\gg 1$, and $w \ll d$. Equation (57) represents a sequence of narrow pulses that occur at every $|y-y'|=nh_0$, where n is a positive integer, and their peak values are weighted by a broad sinc factor. Thus, a high degree of spatial coherence can be achieved at every point pair between the two input color-image transparencies. By applying the Van Cittert–Zernike Theorem of Eq. (39), the corresponding source encoding function is obtained:

$$\gamma(|y|)=\sum_{n=1}\operatorname{rect}\frac{|y-nd|}{w} \tag{58}$$

where w is the slit width, $d=(\lambda f/h_0)$ is the separation between the slits, f is the focal length of the achromatic collimated lens, and N is the total number of slits. Alternatively, Eq. (58) can be written in the form

$$\gamma(|y|)=\sum_{n=1}^{N}\operatorname{rect}\frac{|y-n\lambda f/h_0|}{W} \tag{59}$$

for which we see that the source encoding mask is essentially a fan-shape grating. To obtain lines of rainbow-color spectral light sources for the subtraction operation, we would use a linear extended white-light source with a dispersive phase grating, as illustrated in Fig. 17. For broad-spectral-band image substraction operation, a fan-shape sinusoidal grating is utilized in the Fourier plane, such as

$$G=\tfrac{1}{2}[1+\sin(2\pi\alpha h_0/\lambda f)] \tag{60}$$

The output image irradiance around the optical axis can be shown as

$$I(x,y)=K|s_1(x,y)-s_2(x,y)|^2 \tag{61}$$

Thus, we assert that a color-subtracted image can readily be obtained.

Figures 18a and 18b show a set of black-and-white pictures of input color transparencies. The first picture shows an F-16 fighter plane taking off over a mountainous coastline area, and the second picture shows the same scenic view but without the fighter plane. Figure 18c shows the corresponding subtracted image at the output plane. From this figure, we see that the shape of the F-16 fighter plane can be recognized. Although the resolution of the result is still below the generally acceptable standard, this drawback may be overcome by utilizing higher-quality optics.

C. Complex Signal Detection

One of the most important applications of optical signal processing is complex signal detection. We shall now describe a white-light optical correlator that can exploit the spectral content of the object under observation. In

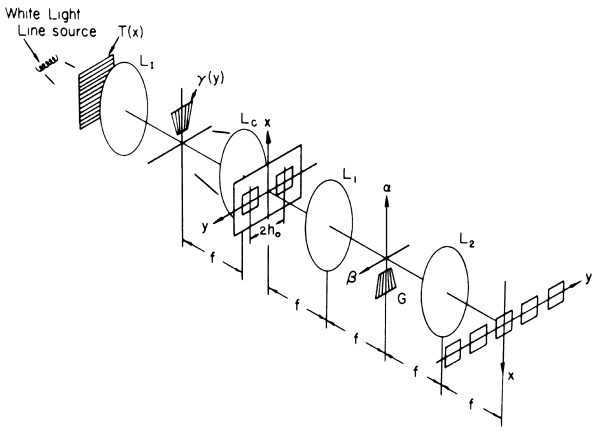


FIGURE 17 A white-light image subtraction system: T(x), phase grating; L_1, imaging lens; L_C, collimated lens; L_1 and L_2, achromatic transform lenses; $\gamma(y)$, fan-shaped source encoding mask; G, fan-shaped diffraction grating.

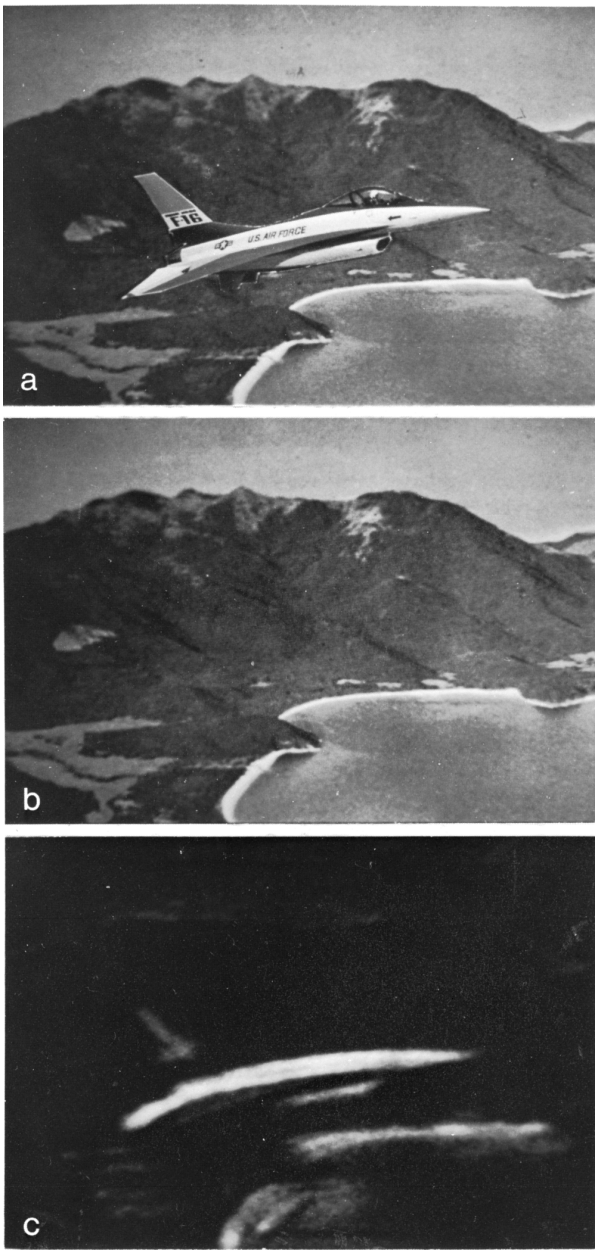


FIGURE 18 Image subtraction. (a) and (b) Black-and-white pictures of input color objects. (c) A black-and-white picture of the output subtracted color image.

other words, the correlator is capable of recognizing multiple color objects of different shapes. The need for color-pattern recognition is very diverse and offers many applications. There are, however, two large categories where color is extremely important: natural color variations, and objects deliberately colored for identification, for example, in robotic vision applications.

We assume that a broad-band matched filter, as described in the following equation, is provided:

$$H(\alpha,\beta) \approx \sum_{\substack{n=-N\\ n\neq 0}}^{N} \left\{ K_1 |S(\alpha,\beta + f\nu_0\lambda_n)|^2 + K_2 |S(\alpha,\beta + f\nu_0\lambda_n)| \times \cos\left[\frac{2\pi h}{\lambda_n f}\alpha + \phi(\alpha,\beta + f\nu_0\lambda_n)\right]\right\} \tag{62}$$

where $S(\alpha,\beta;\lambda)$ is the input object spectrum, ν_0 is the sampling grating frequency, K_s, C_s, and h are arbitrary constants, and λ_n is the main wavelength of the nth spectral-band filter. If this filter is inserted in the Fourier plane of a white-light processor, as depicted in Fig. 4, then the output complex light filter can be shown to be

$$g(x,y;\lambda) = \sum_{n=1}^{N} C_1 s(x,y;\lambda_n) * s(x,y;\lambda_n) * s^*(-x,-y;\lambda_n) + \sum_{n=1}^{N} C_2 s(x,y;\lambda_n) * [s(x+h,y;\lambda_n) + s^*(-x+h,-y;\lambda_n)] \tag{63}$$

where $s(x,y;\lambda_n)$ is the color input object, $*$ denotes the convolution operation, and the superscript * represents the complex conjugate. The corresponding output irradiance is

$$I(x,y) = \int_{\Delta\lambda_n} |g(x,y;\lambda)|^2 \, d\lambda \approx \sum_{n=1}^{N} \Delta\lambda_n \{ K_1 |s(x,y;\lambda_n) * s(x,y;\lambda_n) \circledast s^*(x,y;\lambda_n)|^2 + K_2 |s(x,y;\lambda_n) * s(x+h,y;\lambda_n) + s(x,y;\lambda_n) \circledast s^*(x-h,y;\lambda_n)|^2 \} \tag{64}$$

where $*$ denotes the correlation operation, and K_s are the proportionality constants. Thus, we see that an autocorrelation of the input image is diffracted at $x = h$ at the output plane. Since we utilized a broad-spectral-band white-light source, the spectral content of the image can be exploited.

Figure 19a shows a black-and-white photograph of a color aerial photograph with a (red) guided missile in flight. Figure 19b shows a visible (red) correlation spot obtained at the output plane of the correlator. Thus the white-light correlator has a unique feature for color pattern recognition based on both the spectral content and the shape of and input object.

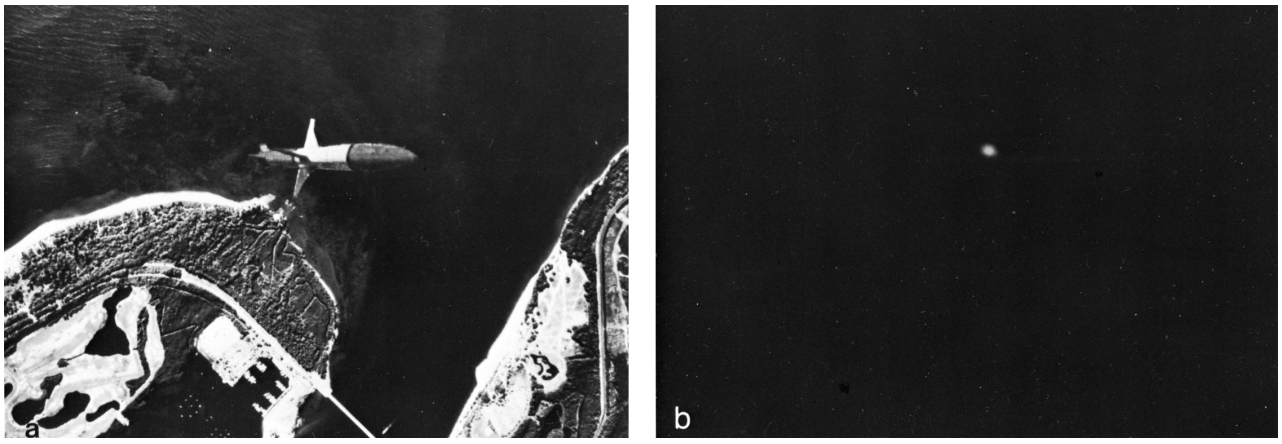


FIGURE 19 Complex signal detection. (a) A black-and-white picture of an input color object. (b) A black-and-white picture of the output correlation spot.

D. Color-Image Retrieval

We shall now describe a spatial encoding technique for color-image retrieval. A color transparency is used as an object to be encoded by sequentially exposing with the primary colors of light, onto a black-and-white film. The encoding takes place by spatially sampling the primary color images of the color transparency with a specific sampling frequency and a predescribed direction onto a monochrome film. In order to avoid the moire fringe pattern, the primary color images are sampled in orthogonal directions. Thus the intensity transmittance of the encoded films can be shown as

$$\begin{aligned} T(x, y) = K\{&T_{\mathrm{r}}(x, y)[1 + \operatorname{sgn}(\cos \omega_{\mathrm{r}} y)] \\ &+ T_{\mathrm{b}}(x, y)[1 + \operatorname{sgn}(\cos \omega_{\mathrm{b}} x)] \\ &+ T_{\mathrm{g}}(x, y)[1 + \operatorname{sgn}(\cos \omega_{\mathrm{g}} x)]\}^{-\gamma} \end{aligned} \tag{65}$$

where K is an appropriate proportionality constant; T_{r}, T_{b}, and T_{g} are the red, blue, and green color-image exposures; ω_{r}, ω_{b}, and ω_{g} are the respective carrier spatial frequencies; (x, y) is the spatial coordinate system of the encoded film; γ is the film gamma; and

$$\operatorname{sgn}(\cos x) \equiv \begin{cases} 1 & \cos x \geq 0 \\ -1 & \cos x \leq 0 \end{cases}$$

To improve the diffraction efficiency of the encoded film, it is advisable to bleach the encoded film to convert it into a phase-type transparency. The amplitude transmittance of the bleached transparency can be written as

$$t(x, y) = \exp[i\phi(x, y)]$$

where

$$\begin{aligned} \phi(x, y) = M\{&T_{\mathrm{r}}(x, y)[1 + \operatorname{sgn}(\cos \omega_{\mathrm{r}} y)] \\ &+ T_{\mathrm{b}}(x, y)[1 + \operatorname{sgn}(\cos \omega_{\mathrm{b}} x)] \\ &+ T_{\mathrm{g}}(x, y)[1 + \operatorname{sgn}(\cos \omega_{\mathrm{g}} x)]\} \end{aligned} \tag{66}$$

and M is an appropriate proportionality constant.

If we place this bleached encoded film in the input plane of the white-light optical processor of Fig. 20, then the first-order complex light distribution for every λ at the spatial frequency plane P_2 can be shown as

$$\begin{aligned} S(\alpha, \beta; \lambda) \approx{}& \hat{T}_{\mathrm{r}}\left(\alpha, \beta \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{r}}\right) + \hat{T}_{\mathrm{b}}\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{b}}, \beta\right) \\ &+ \hat{T}_{\mathrm{g}}\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{g}}, \beta\right) + \hat{T}_{\mathrm{r}}\left(\alpha, \beta \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{r}}\right) \\ &* \hat{T}_{\mathrm{b}}\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{b}}, \beta\right) + \hat{T}_{\mathrm{r}}\left(\alpha, \beta \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{r}}\right) \\ &* \hat{T}_{\mathrm{g}}\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{g}}, \beta\right) + \hat{T}_{\mathrm{b}}\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{b}}, \beta\right) \\ &* \hat{T}_{\mathrm{g}}\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_{\mathrm{g}}, \beta\right) \end{aligned} \tag{67}$$

where $\hat{T}_{\mathrm{r}}$, $\hat{T}_{\mathrm{b}}$ and $\hat{T}_{\mathrm{g}}$ are the Fourier transforms of T_{r}, T_{b}, and T_{g}, respectively, $*$ denotes the convolution operation, and the proportionality constants have been neglected for

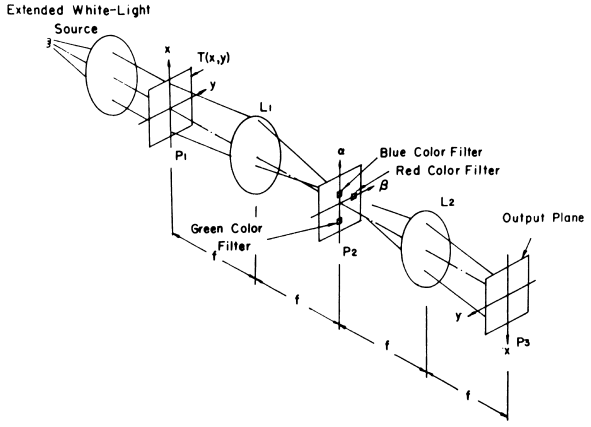


FIGURE 20 A white-light processor for color-image retrieval.

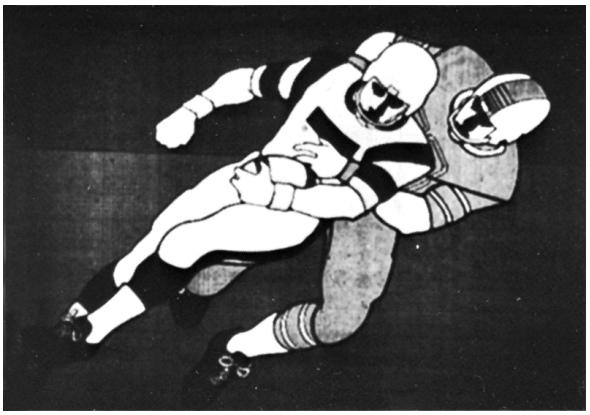


FIGURE 21 A retrieved color image.

simplicity. We note that the last cross-product term would introduce a moire fringe pattern, which can be easily masked out at the Fourier plane. It is apparent that, by proper color-filtering of the smeared Fourier spectra, a true color image can be retrieved at the output image plane. Thus, the output image irradiance would be

$$I(x, y) = T_r^2(x, y) + T_b^2(x, y) + T_g^2(x, y) \tag{68}$$

which is a superposition of three primary encoded color images.

Let us now provide a result of the retrieved color image obtained with this technique, as shown in Fig. 21. As compared with the original color transparency, we can see that the retrieved color is spectacularly faithful and has virtually no color cross-talk. The resolution and contrast are still below the acceptance for wide-spread applications; however, these drawbacks may be overcome by utilizing a more suitable film.

E. Density Pseudo-Color Encoding

Most of the optical images obtained in various scientific applications are gray-level density images—for example, the scanning electron micrographs, multispectral-band aerial photographic images, X-ray transparencies, infrared scanning images, and many others. However, humans can perceive details in color better than in gray levels. In other words, a color-coded image can provide better visual discrimination.

We shall now describe a density pseudo-color encoding technique for monochrome images. We assume that a gray-level transparency (called T_1) is available for pseudo-coloring. By a contact printing process, a negative, and a product (called T_2 and T_3, respectively) image transparency can be made. It is now clear that the spatial encoding onto a monochrome film can take place using the procedure described for color image retrieval. The intensity transmittance of the encoded film can be written as

$$\begin{aligned} T(x, y) = K\{&T_1(x, y)[1 + \operatorname{sgn}(\cos \omega_1 y)] \\ &+ T_2(x, y)[1 + \operatorname{sgn}(\cos \omega_2 x)] \\ &+ T_3(x, y)[1 + \operatorname{sgn}(\cos \omega_3 x)]\}^{-\gamma} \end{aligned} \tag{69}$$

where K is an appropriate proportionality constant. To improve the diffraction efficiency of the encoded film, again we utilize a bleaching process such that the phase encoded transparency becomes

$$t(x, y) = \exp[i\phi(x, y)]$$

where

$$\begin{aligned} \phi(x, y) = M\{&T_1(x, y)[1 + \operatorname{sgn}(\cos \omega_1 y)] \\ &+ T_2(x, y)[1 + \operatorname{sgn}(\cos \omega_2 x)] \\ &+ T_3(x, y)[1 + \operatorname{sgn}(\cos \omega_3 x)]\} \end{aligned} \tag{70}$$

where M is an appropriate proportionality constant. If we insert this encoded phase transparency at the input plane of the white-light processor of Fig. 20, the complex light distribution due to $t(x, y)$ for every λ at the spatial frequency plane is

$$\begin{aligned} S(\alpha, \beta, \lambda) \approx{}& \hat{T}_1\left(\alpha, \beta \pm \frac{\lambda f}{2\pi}\omega_1\right) + \hat{T}_2\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_2, \beta\right) \\ &+ \hat{T}_3\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_3, \beta\right) + \hat{T}_1\left(\alpha, \beta \pm \frac{\lambda f}{2\pi}\omega_1\right) \\ &* \hat{T}_2\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_2, \beta\right) + \hat{T}_1\left(\alpha, \beta \pm \frac{\lambda f}{2\pi}\omega_1\right) \\ &* \hat{T}_3\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_3, \beta\right) + \hat{T}_2\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_2, \beta\right) \\ &* \hat{T}_3\left(\alpha \pm \frac{\lambda f}{2\pi}\omega_3, \beta\right) \end{aligned} \tag{71}$$

where $\hat{T}_1$, $\hat{T}_2$, and $\hat{T}_3$ are the smeared Fourier spectra of the positive, negative, and product images respectively, $*$ denotes the convolution operation, and the proportionality constants have been neglected for simplicity. Again, we see that the last cross product (i.e., the moire fringe pattern) can be avoided by spatial filtering. Needless to say, by proper color filtering of the first-order smeared Fourier spectra, a moire-free pseudo-color encoded image can be obtained at the output plane. The corresponding pseudo-color image irradiance is therefore

$$I(x, y) = T_{1r}^2(x, y) + T_{2b}^2(x, y) + T_{3g}^2(x, y) \tag{72}$$

where T_{1r}^2, T_{2b}^2, and T_{3g}^2 are the red, blue, and green intensity distributions of the three spatially encoded images.

Figure 22 shows a black-and-white picture of a color-coded X-ray image of a human skull. The positive image



FIGURE 22 A black-and-white picture of a density pseudo-color encoded X-ray image.

was encoded in blue, the negative image in red, and the product image in green. By comparing the color-coded image with the original X-ray transparency, we have noted that the soft-tissue objects in the color-coded image can be more easily differentiated by the human eye.

F. Generation of Speech Spectrograms

We shall now describe a technique of generating a multicolor speech spectrogram with a white-light optical signal processor. This technique utilizes a cathode-ray-tube (CRT) scanner to convert a temporal signal to a spatial signal recorded onto a moving photographic film. If this encoded film is transporting over the input optical window of a white-light spectrum analyzer, as shown in Fig. 23, then a speech spectrogram can be generated at the output plane.

We stress that for time to spatial signal conversion, the speed of the film motion is required to satisfy the inequality

$$v \geq \nu / R, \tag{73}$$

where R is the spatial resolution of the film and ν is the highest-frequency content of the time signal.

If the encoded signal format is transporting over the input plane of the white-light processor of Fig. 23, a slanted set of nonoverlapping smeared color spectra in the spatial frequency plane can be observed. It is now clear that if a slanted narrow slit is utilized at the Fourier plane, a frequency color-coded spectrogram can be recorded at the output plane. We further note that, by simply varying the width w of the input optical window, one would obtain the so-called *wide-band* and *narrow-band* spectrograms. In other words, if a broader optical window is utilized, then a narrow-band speech spectrogram, which corresponds to higher spectral resolution, can be obtained. On the other hand, if the optical window is narrower, then a wide-band spectrogram will be obtained. The loss of the spectral resolution due to a narrower optical window would, however, improve the time resolution of the spectrogram. This time–bandwidth relationship is, in fact, the consequence of the well-known Heisenberg's uncertainty relation in quantum mechanics.

Figure 24 shows a black-and-white picture of a frequency color-coded speech spectrogram obtained with the

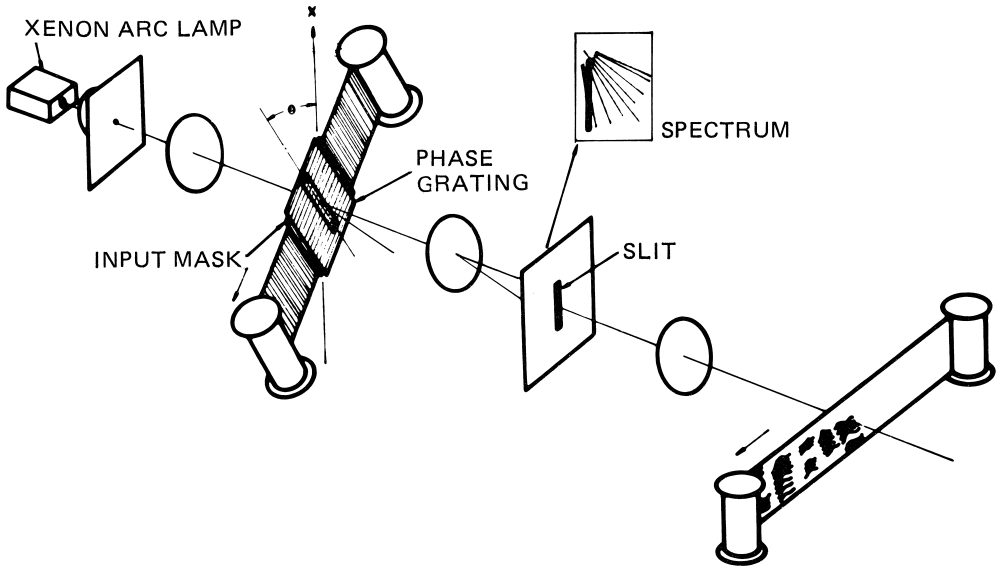


FIGURE 23 A white-light optical sound spectrograph.

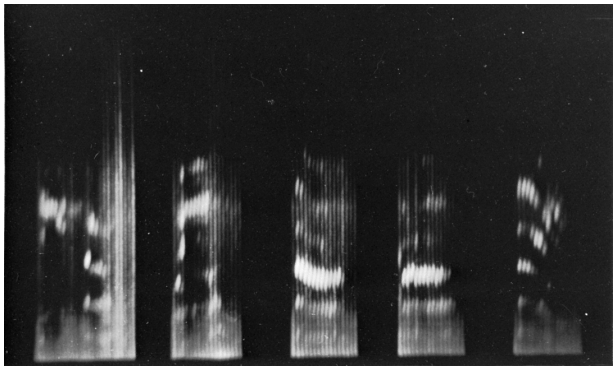


FIGURE 24 A black-and-white picture of a frequency color-coded speech spectrogram.

proposed white-light technique. The frequency content is encoded with red for high frequency, green for intermediate frequency, and blue for low frequency. This speech spectrogram represents a sequence of English words spoken by a male voice. These words are "testing, one, two, three, four." From this speech spectrogram, the format variation can readily be identified. As compared with the electronic and digital counterparts, the optical technique simplified the processing procedure, and the system is rather versatile to operate.

G. White-Light Optical Neural Network

In terms of calculation such as addition, subtraction, multiplication, and division, an electronic computer can compute far more efficiently than a human brain. In terms of pattern recognition, however, a 3-yr-old child can perform better than a super computer. In recent years, scientists have been looking into the possibility of simulating an artificial neuron network for computing, called neural computing.

Two of the most important features of an artificial neural network must be the possibility of massive interconnections and parallel processing operations. In 1985, Farhat and Psaltis suggested using an optical technique to perform the massive interconnection in an optical neural network, in which they used the basic concept of a vector-matrix optical processor.

We have used a video monitor as a programmable interconnection weight matrix (IWM) for an optical neural network implementation. The physical size of the monitor is large, however, and the curvature of the monitor screen poses some problems. To alleviate these two major problems, we developed a white-light optical neural network using pocket-sized liquid crystal televisions (LCTVs).

Generally speaking, an N neurons network requires up to N^2 interconnections. As Hopfield suggested, the state

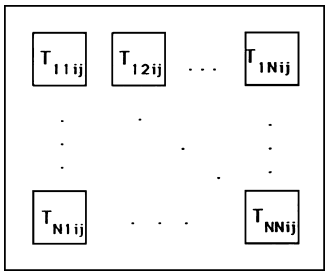


FIGURE 25 Partition of a 4-D matrix T_{1kij} into an array of 2-D submatrices T_{11ij}, T_{12ij}, ... and T_{NNij}.

of each neuron may be expressed by the following iterative equation:

$$V_i(n+1) = f\left(\sum_{j=1}^{N} T_{ij} V_j(n)\right), \qquad i = 1, 2, \ldots, N, \tag{74}$$

where V_i is the state of the ith neuron, n stands for the nth iteration, T_{ij} represents the connection from the jth to the ith neurons, and f represents a nonlinear operator.

It is apparent that the iterative equation can be extended for a 2-D $N \times N$ neurons network, such as:

$$v_{1_k}(n+1) = f\left(\sum_{i=1}^{N}\sum_{j=1}^{N} T_{1_{k_{ij}}} V_{i_j}(n)\right) \tag{75}$$

where v_{1_k} represents the state of the $1k$th neuron in an $N \times N$ space, and $T_{1_{k_{ij}}}$ is a 4-D interconnection weight matrix. Notice that the matrix T can be partitioned into an array of 2-D submatrices and each submatrix is an $N \times N$ size, as shown in Fig. 25.

A white-light optical neural network that we developed is shown in Fig. 26. A Hitachi 5-in. LCTV is used to display the IWM, and a Seiko 2.7-in. LCTV is used to display the input patterns. The resolutions of these LCTVs are 256×420 pixels and 220×330 pixels, respectively.

The lenslet array, consisting of 8×8 lenses (equivalent to 64 neurons), provides the interconnections between the

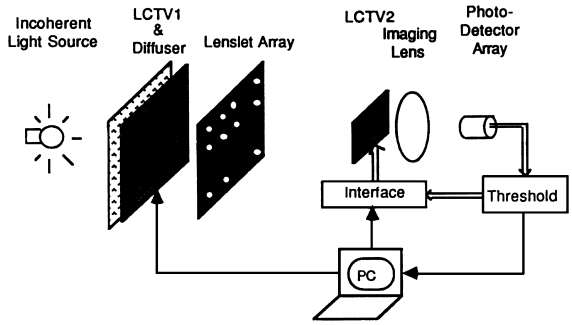


FIGURE 26 Architecture of a white light optical neural network.

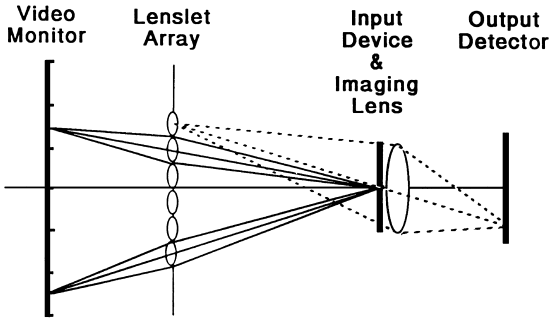


FIGURE 27 A 1-D representation of optical interconnections.

IWM and the input pattern. Each lens of the lenslet array images each of the IWM submatrices onto the input LCTV2 to establish the interconnections shown in Fig. 27. Thus, the IWM submatrices can be superimposed, added and then multiplied (or interconnected) by the input pattern. This represents the connective part of Eq. (75), i.e.,

$$\sum_{i=1}^{N}\sum_{j=1}^{N} T_{1_{k_{ij}}} V_{i_j}(n). \tag{76}$$

The output result, behind the LCTV2, is then imaged onto a charge-coupled detector (CCD) camera. The signals are then sent to a thresholding circuit, and the final results can be fed back to the LCTV2 for second iteration. Notice that the data flow in the optical system is primarily controlled by a personal computer (PC), for which the IWM and the input pattern can be programmed. Thus, the white-light optical architecture is a programmable and adaptive neural network.

The white-light optical neural network is depicted in Fig. 28. From left to right: a xenon arc lamp is used as the light source, a Hitachi color LCTV, model C5-LC1, is used to display the IWM. An 8×8 lenslet array of planoconvex lenses (6 mm diameter and 36 mm focal length), and sandwiched by two optical flat glass plates, is used to establish the optical interconnection between LCTV1 and LCTV2. A Seiko color LCTV, model LVD-202, is used for input pattern generation. A set of imaging lenses (two identical planoconvex lenses of 100 mm diameter and 125 mm focal length, sealed face to face in a mount, with an effective 70 mm focal length) is used to image the output result from LCTV2 onto the CCD camera. Notice that this arrangement of planoconvex lenses has the ability to correct lens distortion. At the rightmost end of the system, a SONY CCD camera, model XC-77, is used as an output array detector.

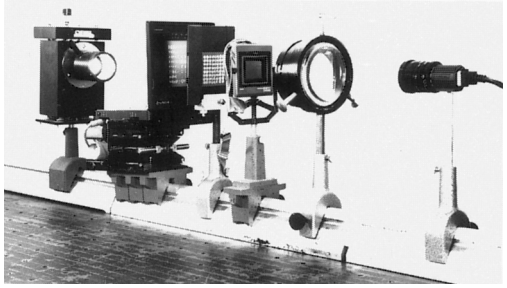


FIGURE 28 A white-light optical neural network.

The size of the optical neural network is measured to be about $65 \times 12 \times 12$ cm^3. Notice that, if a smaller size LCTV is used for IWM generation (for example a 7 cm Epson color LCTV), then the system size can be further reduced to approximately $20 \times 7 \times 7$ cm^3. In this case, the diameter of the lenses for a 16×20 lenset array is within the submillimeter range. We note that this lenslet array can be synthesized by binary optics or holographic techniques.

Since the LCTV1 has a resolution of 256×420 pixels, a maximum number of $16 \times 20 = 320$ neurons can actually be built. The processing speed is limited by the devices within the system, which have an addressing speed of 30 frames/sec. Thus the operational speed of the optical neural network with 8×8 neurons is about $30 \times (8 \times 8)^2 = 1.2 \times 10^5$ interconnection operations/sec. However, for a 16×20 neural network, the operational speed can be as high as $30 \times (16 \times 20)^2 = 3 \times 10^6$ interconnection operations/sec.

We have used the Hopfield, Orthogonal Protection (OP), and Interpattern Association (IPA) models for the generation of the IWMs. One of the IPA experimental demonstrations is shown in Fig. 29. Notice that the capital English letters, shown in Fig. 29(a), are used as reference patterns,

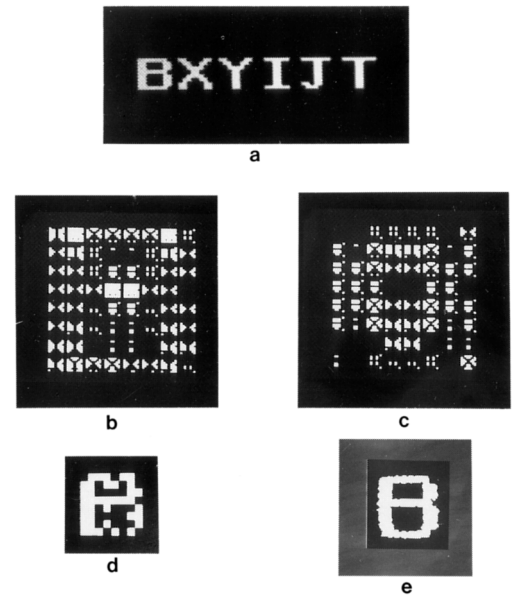


FIGURE 29 Experimental demonstration: (a) training patterns; (b) and (c) positive and negative parts of the IWM using IPA model; (d) input pattern embedded in 30% noise; (e) reconstructed result.

and each letter is displayed in an 8×8 pixel array matrix. The positive and negative parts of the IWM generated by the IPA model are shown in Figs. 29(b) and 29(c), respectively. There are, however, ways to represent the negative values in the IWM, by adding a bias into the IWM, or by using an area-encoding method. For simplicity, the positive and negative parts of the IWM are sequential in our experimental demonstrations. An input pattern B, embedded in 30% random noise, is presented at the LCTV2, as shown in Fig. 29(d). The input pattern B is then multiplied by the submatrices of IWM, by the set of lenslet array. The positive and negative parts of the output results are captured by the CCD camera. The subtraction and thresholding operations are performed by the PC, for which the output result is shown in Fig. 29(e). Thus, we can see that a recognizable B is reconstructed.

VIII. CONCLUDING REMARKS

We have presented the basic concept of optical processing with partially coherent light. We have reviewed several partially coherent optical processing techniques. A mathematical formulation for a partially coherent processor with a white-light source was obtained. The performance of a partially coherent processor can be, in principle, evaluated with this generalized formulation. The principle of source encoding and image sampling as applied to the coherence requirement of a partially coherent processor has also been discussed. The objective of utilizing a source encoding technique for a partially coherent processor is to alleviate the stringent spatial coherence requirement so that a practical extended source can be used. The image sampling is to improve the degree of temporal coherence in Fourier plane, so that the image can be processed in complex light field for a broadband light source. Since the partially coherent processing can be carried out by a white-light source, the technique is very suitable for color-image processing. We have also, in this chapter, demonstrated several applications. We have shown that the partially coherent processing system is capable of performing image deblurring, image subtraction, complex signal detection, color image retrieval, pseudo-color encoding, speech spectrogram generation, and white-light optical neural network. We have noted that image processing with white-light is generally very economical and that the technique is rather easy to implement, in contrast with its coherent counterparts.

While the electronic digital signal processor offers flexibility, the partially coherent technique offers the advantages of higher resolution, wavelength diversity, simplicity, and reduced cost. With the current advances of electrooptical spatial light modulators, and optical-digital interface devices, we would expect a gradual merging of the optical and digital techniques; namely, hybrid or microcomputer-based optical signal processing. By utilizing the strengths and merits of both processing techniques, we would expect that many fruitful results will be reported. Furthermore, we believe that partially coherent processing is at the threshold of widespread application. We hope that this chapter will serve as an introductory entry toward the vast application of partially coherent processing.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICIAL NEURAL NETWORKS • COLOR SCIENCE • HOLOGRAPHY • IMAGE PROCESSING • NONLINEAR OPTICAL PROCESSES

BIBLIOGRAPHY

Chang, B. J., and Winick, K. (1980). Silver-halide gelatin holograms. *SPIE* **215,** 1972.

Chao, T. H., Zhuang, S. L., Mao, S. Z., and Yu, F. T. S. (1983). Broad spectral band color image deblurring. *Appl. Opt.* **22,** 1439.

Chavel, P., and Lowenthal, S. (1978). Noise and coherence in optical image processing. II. Noise fluctuations. *J. Opt. Soc. Am.* **68,** 721.

Das, P. K. (1991). "Optical Signal Processing: Fundamentals," Springer-Verlag.

Farhat, N. H., Psaltis, D., Prata, A., and Paek, E. (1985). Optical implementation of the Hopfield model. *Appl. Opt.* **24**(10), 1469.

Hopfield, J. J. (1982). Neural network and physical system with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **79,** 2554.

Leith, E. N. (1980). Image deblurring using diffraction grating. *Opt. Lett.* **5,** 70.

Leith, E. N., and Roth, J. (1977). White-light optical processing and holography. *Appl. Opt.* **16,** 2565.

Lu, T., Xu, X., Wu, S., and Yu, F. T. S. (1990). Neural network model using interpattern association. *Appl. Opt.* **29,** 284.

Madsen, C. K., and Zhao, J. H. (1999). "Optical Filter Design and Analysis: A Signal Processing Approach," Wiley-Interscience, New York.

Morris, G. M., and George, N. (1980). Matched filtering using band-limited illumination. *Opt. Lett.* **5,** 202.

Rumelhart, D. E., and McClelland, J. L., eds. (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition," Vols. 1 and 2. MIT Press, Cambridge, Mass.

Vanderlugt, A. (1992). "Optical Signal Processing," Wiley-Interscience, New York.

Wu, S., Lu, T., Xu, X., and Yu, F. T. S. (1989). An adoptive optical neural network using a high resolution video monitor. *Micro. Opt. Tech. Lett.* **2,** 252.

Yu, F. T. S. (1976). "Optics and Information Theory," Wiley-Interscience, New York.

Yu, F. T. S. (1985). "White Light Optical Signal Processing," Wiley-Interscience, New York.

Yu, F. T. S., and Hsu, F. K. (1985). Complex signal detection with broad band white-light processing. *Appl. Opt.* **24,** 2135.

Yu, F. T. S., and Wu, S. T. (1982). Color image subtraction with extended incoherent sources. *J. Opt.* **13,** 183.

Yu, F. T. S., Zhuang, S. L., and Wu, S. T. (1982). Source encoding for partially coherent optical processing. *Appl. Phys.* **B27,** 99.

Yu, F. T. S., Chen, X. X., and Chao, T. H. (1984). Density pseudocolor encoding with three primary colors. *J. Opt.* **15,** 55.

Yu, F. T. S., Lin, T. N., and Xu, K. B. (1985). White-light optical speech spectrogram generation. *Appl. Opt.* **26,** 836.

Yu, F. T. S., Chen, X. X., and Zhuang, S. L. (1985). Progress report on archival storage of color films utilizing a white-light processing technique. *J. Opt.* **16,** 59.

Yu, F. T. S., Zheng, L. N., and Hsu, F. K. (1985). Noise measurement of a white-light optical signal processor. *Appl. Opt.* **24,** 173.

Yu, F. T. S., Lu, T., Yang, X., and Gregory, D. A. (1990). Optical neural network using pocket size liquid crystal televisions. *Opt. Lett.*

Yu, F. T. S., and Jutamulia, S. (1996). "Optical Storage and Retrieval: Memory, Neural Networks, and Fractals," Dekker, New York.



Radar

Nadav Levanon
Tel Aviv University

I. History and Fundamentals
II. The Radar Equation
III. Radar Cross Section
IV. Radar Detection
V. Radar Propagation
VI. Surface Backscattering (Clutter)
VII. Pulse Compression
VIII. Doppler Filtering
IX. Antennas and Tracking
X. Special Radars

GLOSSARY

Autocorrelation For a function $f(t)$, the integral from $-\infty$ to ∞ of $f(t)f(t-\tau)$, where τ is a time-delay parameter.
Clutter Radar reflection from ground or sea surface.
Correlation Measure of the tendency of two (or more) random variables to vary together.
Decibel (dB) Unit for describing the ratio of two power levels P_1 and P_2, measured by $10\log_{10}(P_2/P_1)$.
Impulse response Output function of a linear system when the input is an impulse of infinitesimal duration and unit area; defines the system.
Multipath Combined propagation through the line of sight and through indirect reflections.
Polarization Orientation of the electric field of electromagnetic wave.
Sidelobes Undesired peaks off the main lobe of an antenna radiation pattern or of a system response.
Specular Mirror-like reflection of radar signals.

RADAR is the acronym for RAdio Direction And Range. It describes a technique or an instrument which determines the location of a target by transmitting an electromagnetic signal toward it and measuring the delay and bearing of the echo.

I. HISTORY AND FUNDAMENTALS

The acronym *radar* was coined in Great Britain in the pre-World War II years of intensive pioneering research effort aimed at providing Britain with early detection of enemy aircrafts. Similar development work was simultaneously

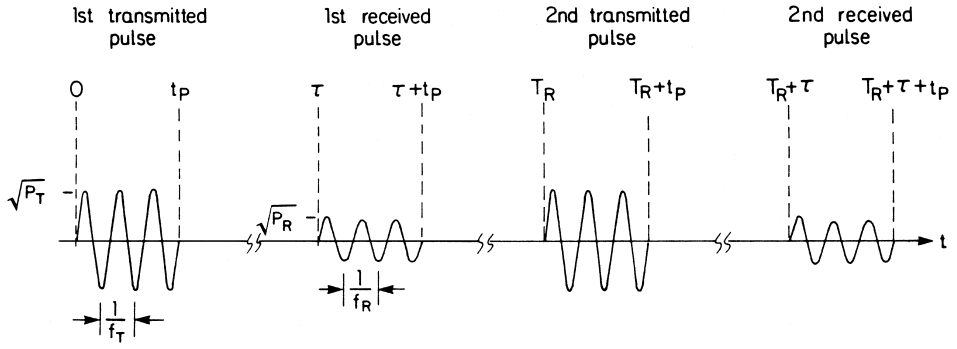


FIGURE 1 Waveforms in a typical pulse radar.

conducted in the United States, Germany, and other countries. The early activity in the United States originated from research aimed at determining the height of the ionosphere. It was then split into two efforts: providing projectiles with the ability to sense proximity to air and surface targets—the proximity fuze; and full-fledged radars in microwave frequencies utilizing powerful transmitters and narrow beam antennas.

Today, radar has found a large variety of military and civil applications. Some of the civil applications are air traffic control, aircraft and ship navigation and safety, remote sensing, and law enforcement. Examples of military applications are surveillance, navigation, weapon control, fuzing, artillery location, and reconnaissance.

In its basic concept, a radar instrument emits a narrow electromagnetic pulse of very short wavelength and measures the delay until the pulse, reflected from a target, is received back. A simple relationship exists between the range R and the delay τ,

$$\tau = \frac{2R}{C_p}, \tag{1}$$

where C_p is the propagation velocity of electromagnetic waves (the speed of light). In vacuum $C_p = C =$ 299792500 m/sec. In the atmosphere C_p is slightly slower and changes with height. The effect of the profile of C_p with height will be discussed in Section V.

The remaining parameters, besides range, necessary in order to determine the position of a target relative to the radar are the two bearing parameters—azimuth and elevation. These can be measured by using a steerable narrow beam antenna for transmitting and receiving the radar signal. The direction from which maximum signal is received back is basically the direction to the target. Target bearing accuracy improves with decreasing antenna beamwidth. On the other hand, the antenna beamwidth is inversely proportional to the antenna aperture A. If θ_H and θ_V are the horizontal and vertical beamwidths in radians (1 rad = 57.3°), then the following approximate relationship holds:

$$\theta_H \theta_V \approx \frac{\lambda^2}{A}. \tag{2}$$

The above relationship demonstrates that in order to achieve a narrow beamwidth with a manageable sized antenna, it is necessary to use a very short wavelength λ. This explains the importance to radar development of the invention in 1939 of a powerful microwave oscillator—the magnetron.

Early radars measured range and direction only. Modern radars can also measure the target's range rate, that is, the radial velocity component. This can be accomplished by measuring the Doppler frequency shift of the reflected signal. When (Fig. 1) the transmitted radar signal is a pulse modulating a carrier frequency f_T, the signal reflected from a point target with radial velocity $\dot{R}$ will be a similar pulse, modulating a shifted carrier frequency f_R. The difference between the two frequencies is the Doppler shift

$$f_R - f_T \triangleq f_D = -\frac{2\dot{R}}{\lambda}, \tag{3}$$

where $\lambda \triangleq \frac{C_p}{f_T}$.

The ratio between the Doppler frequency and the transmitted frequency is quite small. It is only twice the ratio between the target's radial velocity and the speed of light. In order to be able to sense this small relative difference, the transmitted carrier frequency must be very stable. There should be no frequency or phase modulations of that carrier that are unaccounted for. A radar that performs in that manner is called a coherent radar. Magnetron oscillators do not maintain coherency and must be replaced by a low-power stable oscillator followed by a high-power amplifier like a klystron or a traveling wave tube (TWT). The stable oscillator is also used as a reference in the radar receiver.

The ability to detect the presence of a target and to estimate its parameters—range, range rate, and bearing—is determined by the signal-to-noise ratio (SNR). The "signal" is the radar signal reflected from the target. The term "noise" refers to both the thermal noise, inherent in every electronic circuit, and also the undesired signals,

such as reflections from ground and other targets and intentional interferences known as electronic counter measures (ECM).

The most prominent difference between radar and communication is the fact that the radar receiver knows exactly what was the transmitted signal. Such knowledge allows the receiver to match its response to the known signal, in order to yield maximum performances when that signal is received, while suppressing the effects of noise and clutter. The desired performances are: (a) achieving high SNR in order to improve the probability of detection and the accuracy of estimating the position and velocity of the target, and (b) achieving high resolution in order to be able to separate targets, including separation between targets and ground clutter.

II. THE RADAR EQUATION

The power received back from a target in free space can be predicted using simple geometrical considerations. In free space the electromagnetic power transmitted by an isotropic antenna (antenna that radiates equally to all directions) is spread evenly on the envelope of a sphere. A target of area σ, at range R, intercepts power which is proportional to the ratio of the target area over the area of a sphere of radii R. If the target reflects the intercepted power isotropically, then the power received back by the radar antenna is similarly given by the reflected power times the ratio between the antenna area A and the envelope of the same sphere of radii R. Clearly, the radar antenna does not radiate isotropically but concentrates the power into a narrow beam. This increases the power density in the vicinity of the target by the antenna power gain G. Furthermore, the effective receiving area of the antenna in that direction increases with the gain

$$A = \frac{G\lambda^2}{4\pi}. \tag{4}$$

All these considerations lead to an equation for the received instantaneous power P_R as a function of the transmitted instantaneous power P_T,

$$P_R = \frac{P_T G^2 \lambda^2 \sigma}{(4\pi)^3 R^4}. \tag{5}$$

The thermal noise generated within the receiver can be replaced by an equivalent input noise of constant power-spectral-density N_0 (white noise), given by

$$N_0 = FKT_E, \tag{6}$$

where F is the receiver noise figure (a dimensionless quantity related to the quality of the receiver front end, $F > 1$), K is Boltzmann's constant (1.38×10^{-23} W sec/K), and T_E is the temperature in degrees Kelvin. In order to obtain the noise power from the noise power-spectral-density, the frequency response of the radar receiver must be known. As pointed out earlier, an optimal receiver matches its response to the transmitted waveform. If the receiver is matched to the signal during an illumination time T_I, then the SNR at the end of T_I will be

$$\text{SNR} = \frac{P_{AVE} T_I G^2 \lambda^2 \sigma}{(4\pi)^3 R^4 N_0}, \tag{7}$$

where P_{AVE} is the average transmitted power during T_I. Referring to Fig. 1, for example, the average power is related to the instantaneous power during the pulse transmission P_T as

$$P_{AVE} = \frac{t_p}{T_R} P_T. \tag{8}$$

Continuing with that example, if the illumination time extends over M pulses, then $T_I = MT_R$, yielding

$$P_{AVE} T_I = M P_T t_p. \tag{9}$$

Matching the receiver to a train of M pulses is not easily implemented. The receiver must know the initial phase of each pulse; it must process the received pulses coherently. A receiver that is not perfectly matched to the received signal produces less than optimal SNR. This effect can be introduced into the radar equation by adding a loss term.

III. RADAR CROSS SECTION

When constructing the radar equation it was assumed that the target reflects isotropically. Most targets do not behave so ideally. To circumvent this problem, the original target is replaced by an imaginary isotropic target of area σ that would have reflected back to the radar the same power as the original target. The σ area of a target is called its radar cross section (RCS). Usually, the physical area of a real target will be quite different from its σ. Furthermore, most targets exhibit different σ's at different aspect angles and at different radio frequencies.

A. Ideal Bodies

The RCS of ideal bodies can be calculated analytically. The most perfect object is the sphere. The RCS of a sphere is a complicated function of its radius a. The asymptotes are

$$\sigma = \begin{cases} \pi a^2, & \lambda \ll a \\ \pi a^2 9 \left(\dfrac{2\pi a}{\lambda}\right)^4, & a \ll \lambda. \end{cases} \tag{10}$$

The RCS of a sphere is inherently independent of the aspect angle. Furthermore, σ of a large sphere is independent of the wavelength. These two properties make spheres ideal bodies for the calibration of radar instruments.

For a square flat plate of side length a, the RCS is strongly dependent on the angle of incidence θ (measured from the normal):

$$\sigma = \frac{4\pi a^4}{\lambda^2}\left[\frac{\sin(ka\sin\theta)}{ka\sin\theta}\right]^2, \qquad k = \frac{2\pi}{\lambda}. \tag{11}$$

Note that when a large plate ($a \gg \lambda$) is perpendicular to the range vector ($\theta = 0$), its RCS can be much larger than the area a^2, but it drops rapidly off the normal, for example, when $a = 10\lambda$, $\sigma = 1257a^2$ at $\theta = 0^\circ$, but $\sigma = 0$ at $\theta = 2.9^\circ$.

A target that maintains the large RCS of a flat plate without being so sensitive to the angle of incidence is the *corner reflector*. The RCS of the square trihedral has a maximum at the symmetry angle of 54.7°,

$$\sigma_{\max} = \frac{12\pi a^4}{\lambda^2}, \tag{12}$$

and drops to half power (−3 dB) at about 20° off that angle. Eight such trihedrals—the retroflector (Fig. 2)—provide spherical coverage. The retroflector is used in situations when it is desired to enhance the radar detectability, for example, by small boats.

B. Extended Targets

Large practical targets, such as aircraft, boats, cars, or a patch of land or sea, are constructed from multiple scatterers, each with its own σ and range. Their total RCS is obtained from vectorial addition of the individual reflections. A range difference of $\lambda/4$ causes a 180° phase shift and therefore a subtraction, while a range difference of $\lambda/2$ causes in-phase addition. This leads to a rather complex angular dependence of the RCS of practical targets. When such a target is moving, the aspect angle changes, causing large fluctuations of σ. Fluctuations in σ are also observed in nonrigid targets such as the sea surface or wind-affected vegetation. For this kind of target, the RCS becomes a random variable, with a probability density function (PDF) and a correlation function of time. The PDF describes the range of values that σ may take and their probability of occurrence. The correlation describes the rate of change of σ with time.

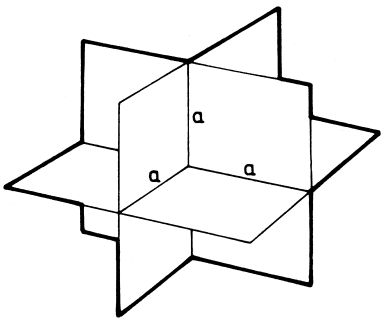


FIGURE 2 Retroflector.

Target fluctuations affect the detection performances of the radar receiver. The receiver deals with the received signal amplitude (voltage), while σ is linearly related (see the radar equation) to the received power. Because the signal power is related to the square of its amplitude x, it is useful to make the transformation $\sigma = x^2/2$.

$$\sigma = \frac{x^2}{2} \tag{13}$$

Many different PDFs of the amplitude of radar-reflected signals are known. The most common one is the Rayleigh PDF

$$p_1(x) = \frac{x}{x_0^2}\exp\left(\frac{-x^2}{2x_0^2}\right): \; 0 \le x. \tag{14}$$

The Rayleigh PDF is predicted theoretically whenever the reflection comes from many scatterers of approximately the same size.

When a train of pulses of equal amplitude, drawn randomly from a single Rayleigh PDF, is reflected from a target, that target is known as a Swerling I target. A train of pulses in which each one of the pulse amplitudes is independent but identically distributed according to the same Rayleigh PDF is defined as coming from a Swerling II target. It can also be stated that the pulses reflected from a Swerling I target are fully correlated, while those that come from a Swerling II target are uncorrelated. Many radar fluctuating targets cannot be described by the simple Rayleigh PDF. Among the many other applicable distributions are the Weibull and log-normal distributions.

In the design of modern fighter aircraft and ships, an effort is made to reduce the vehicle RCS. This is done by shaping the body of the vehicle to avoid corners and large plates facing the prospective radar and by using radiation-absorbing materials.

IV. RADAR DETECTION

Radar detection is the process of distinguishing between targets and noise. It is usually performed by a threshold decision circuit which follows a bandpass filter and an envelope detector. The signal at that point is often called the video signal. A typical video signal is shown in Fig. 3. The decision circuit compares the video signal with a predetermined threshold. Whenever the signal surpasses the

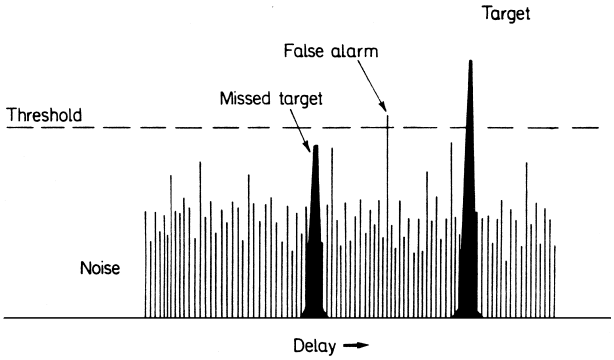


FIGURE 3 Typical radar video signal.

threshold, the existence of a target is assumed at the corresponding delay. The threshold in Fig. 3 happens to be above all but the highest noise peak. That noise peak is mistaken for a target and generates a false alarm. The one target that is too weak to be detected demonstrates that the probability of detection (PD) is smaller than the ideal unit value. Lowering the threshold would have increased the PD, at a cost of also increasing the probability of false alarm (PFA). On the other hand, if the SNR were higher, which implies higher target peaks, the PD would be increased. The threefold relationship between PD, PFA, and SNR is a function of the noise and signal statistics. The relationship is usually too complex to be expressed analytically. An exception is the case when the signal is reflected from a Rayleigh fluctuating target (Swerling I or II) and the noise is band-limited thermal noise. In that case the threefold relationship is given by

$$\mathrm{PD} = \mathrm{PFA}^{\frac{1}{1+\mathrm{SNR}}} \quad \text{or} \quad \mathrm{SNR} = \frac{\log(\mathrm{PFA}/\mathrm{PD})}{\log(\mathrm{PD})}. \qquad (15)$$

A graphic representation of this relationship is given in Fig. 4. As Fig. 4 demonstrates, the requirement for high PD (e.g., PD = 0.95) and low PFA (e.g., PFA = 10^{-5}) can be met with an SNR of 23.5 dB (224). Such a high SNR may be difficult to achieve. The alternative is to base the decision on more than one pulse.

A. Pulse Integration

Integration of N pulses involves summing N samples of the received signal taken at a fixed delay after each one of N transmitted pulses. The fixed delay assures that all the samples in the sum are coming from the same range. If there is a target at that range, the output will be a sum of N target signal samples. Otherwise, the output will be the sum of N noise samples.

Integration of N pulses before making a detection decision allows each pulse to be of smaller level. The largest improvement can be achieved in a coherent radar, where

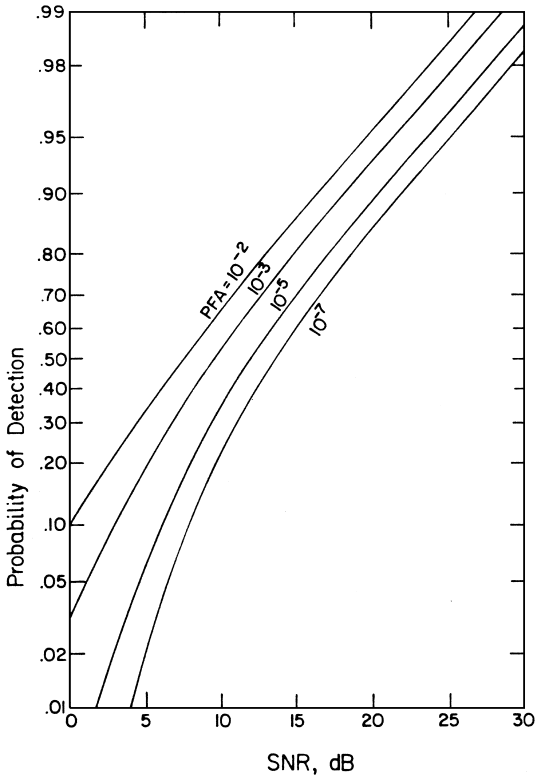


FIGURE 4 Detection performances of a Rayleigh fluctuating signal.

both the amplitude and the phase of each sample are measured. In a coherent radar the sum can be performed vectorially. Because the signal samples are all of the same phase, while the noise samples are of random phase, the summation process favors the signal samples by a factor of N. In other words, the signal power of each one of the received pulses can be $1/N$ the power of a single pulse, in order to get the same detection performances. This kind of integration is known as *coherent integration*. It is possible only in coherent radars and only if the target and the propagation medium do not destroy the coherency. Since envelope detection eliminates the phase information, the coherent integration must be performed before envelope detection, when the signal still modulates a carrier frequency.

Noncoherent integration, also known as post-detection integration, utilizes pulse amplitude only. The improvement in SNR depends strongly on the target statistics and correlation. For example, the sum of N samples taken from the Swerling II target plus noise is generally higher than when the target is Swerling I. Figure 5 presents two PDFs of the sum of 10 signal-plus-noise samples (SNR = 10), where the signal is either a Swerling I or Swerling II type. The PDF of the sum of 10 noise samples is also plotted (divided by 5). In order to guarantee high PD and low PFA,

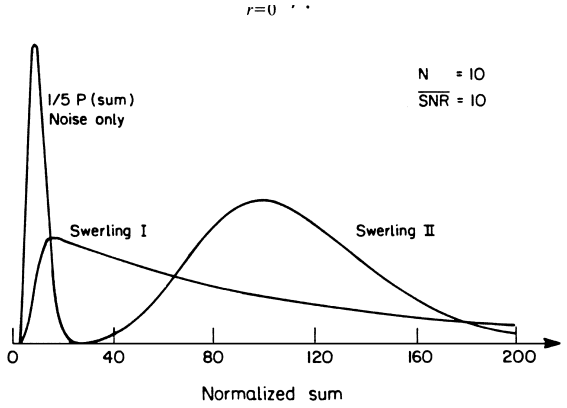


FIGURE 5 PDFs of the sum of 10 samples.

the PDF of the noise samples should not overlap the PDF of the signal-plus-noise samples. For the case represented in Fig. 5, this is accomplished only when the signal comes from a Swerling II target.

For a Swerling II signal in narrow band thermal noise, and when the envelope detector is linearly proportional to the video power (Fig. 6a), the threefold relationship between PFA, PD, and SNR is given by two equations:

$$\mathrm{PFA} = \exp(-T)\sum_{r=0}^{N-1}\frac{T^r}{r!} \tag{16}$$

and

$$\mathrm{PD} = \exp(-TD)\sum_{r=0}^{N-1}\frac{(TD)^r}{r!}, \quad D = \frac{1}{1+\mathrm{SNR}}, \tag{17}$$

where the temporary variable T is the normalized threshold. In order to get the same PD = 0.95 and PFA = 10^{-5} when 10 Swerling II pulses are integrated, the average SNR of each pulse needs to be only 6.5 dB. This is an improvement of 17 dB compared to the required SNR when the detection decision is based on a single pulse.

The integration algorithm discussed above requires analog addition. An integration algorithm that avoids analog addition is the *binary integrator* shown in Fig. 6b. Here, each one of the N samples is compared to the analog threshold. The number of threshold crossing H is counted and compared to a binary threshold M. A target is declared when $H \geq M$. Such an integrator is also called an "M out of N" binary integrator. Besides its simpler digital implementation, the binary integrator exhibits some immunity against strong interferences. However, binary integration is not as efficient as analog integration. For a Swerling II target, PD = 0.95, PFA = 10^{-5}, and $N = 10$, the best binary integrator ($M = 3$) will require an SNR = 8.5 dB. This is 2 dB more than the SNR required by the analog integrator.

B. Constant False Alarm Rate (CFAR)

In a fixed-threshold detector, the PFA can be extremely sensitive to the noise mean power and its statistics. In constant false alarm rate (CFAR) detectors the noise (or background) mean power is continuously estimated, and the threshold adapts itself according to this estimate, in order to maintain a predetermined PFA. An important CFAR algorithm is the cell averaging (CA) CFAR, in which the mean background level is estimated by averaging the signal level in M neighboring range cells. The adaptive threshold is set by multiplying the mean background level by a coefficient α. When the background is a band-limited thermal noise, whose envelope is Rayleigh distributed, the CA-CFAR detector yields the following dependence of the

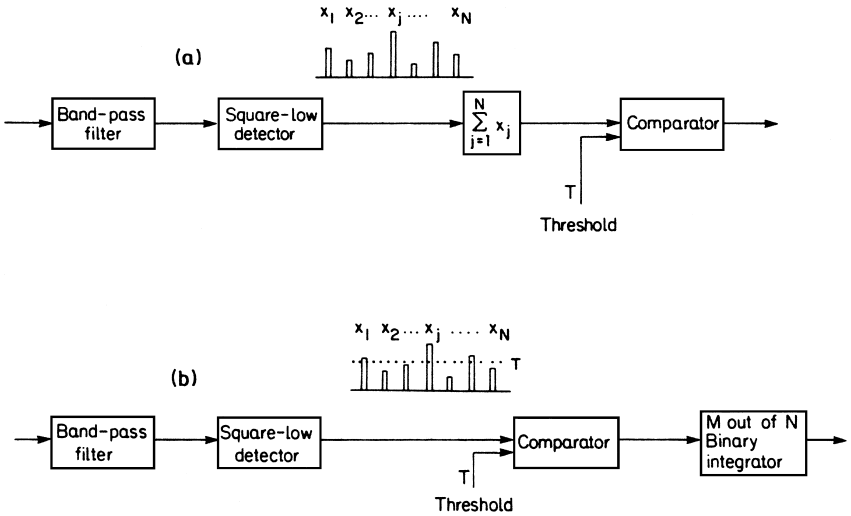


FIGURE 6 Pulse integration: (a) analog, (b) binary.

PFA on the coefficient α and the number M of reference range cells:

$$\mathrm{PFA} = \left(1 + \frac{\alpha}{M}\right)^{-M}. \tag{18}$$

When the range cell under test contains signal from a Swerling I or II fluctuating target, with an average SNR over the background noise, then the threefold relationship between PFA, PD, and SNR becomes

$$\mathrm{SNR} = \frac{(\mathrm{PD}/\mathrm{PFA})^{1/M} - 1}{1 - \mathrm{PD}^{1/M}}. \tag{19}$$

Comparing the required SNR in the CFAR detector (Eq. 19) with the required SNR in the fixed-threshold detector (Eq. 15) shows that the CFAR detector requires higher SNR. The ratio between the two SNRs is called the *CFAR loss*. For the situation discussed above, the loss is approximately

$$\text{CFAR loss} \approx \mathrm{PFA}^{-1/2M}. \tag{20}$$

For the case of $\mathrm{PFA} = 10^{-6}$ and $M = 16$, the CFAR loss is 2 dB.

Another drawback of CFAR detectors is their sensitivity to the statistics of the background cells. An extreme example is the appearance of another target in one of the reference range cells. This scenario will result in an unacceptably high threshold and therefore a drastic reduction in the PD. There are several CFAR algorithms that avoid this problem by censoring the highest reference cells.

V. RADAR PROPAGATION

The radar equation assumes propagation in free space. In a more typical environment, both the atmosphere and the earth surface are present and strongly affect radar propagation.

A. Atmospheric Refraction

The effect of a typical atmosphere is to refract the radar ray downward. Refraction happens because the propagation velocity increases with height. The propagation velocity is a function of the atmospheric parameters

$$C_\mathrm{p} \approx C\left[1 - \frac{77.6}{T}\left(P + 4870\frac{e}{T}\right)10^{-6}\right], \tag{21}$$

where T is the air temperature in degrees Kelvin, P is the air pressure in millibars, and e is the water vapor pressure in millibars. For typical values near sea level height ($h = 0$),

$$C_\mathrm{p} \approx C(1 - 0.000319). \tag{22}$$

Propagation velocity increases with height until about 15 km, above which it is practical to assume $C_\mathrm{p} \approx C$. Maximum downward ray bending occurs when the target is near the horizon. In a standard atmosphere, when a radar near sea surface level measures the target return at an elevation of zero degrees, the target is actually about half a degree below the horizon.

Below a height of 5 km, and under standard atmospheric conditions, ray bending can be predicted by using the "4/3 earth radius" model. In this model, the earth radius is assumed to be 4/3 of its true value, and there is no atmospheric refraction.

The downward bending has the effect of extending the electromagnetic horizon beyond the geometric horizon. When the 4/3 earth radius model applies, the distance to the electromagnetic horizon equals the distance to the geometric horizon of an imaginary earth whose radius is 4/3 the earth radius. Thus, if h_R is the height of the radar antenna, and h_T is the height of the target, the maximum range between antenna and target is

$$R_\mathrm{max} = \sqrt{\frac{8a_\mathrm{E}}{3}}(\sqrt{h_\mathrm{R}} + \sqrt{h_\mathrm{T}}), \quad a_\mathrm{E} = 6370 \text{ km}. \tag{23}$$

Anomalous atmospheric conditions can extend the maximum range much beyond what is predicted by Eq. (23). This phenomenon happens mostly over the sea, where the temperature and moisture gradients can cause a layer with a sharp decrease in the velocity of propagation. That layer and the surface create a duct within which the radar ray can travel to great distances.

B. Earth Forward Reflection (Multipath)

The particular situation of targets at low elevation above the earth surface is further complicated by the reflection from the earth surface. The earth surface, rough as it may be, behaves like a mirror when the grazing angle α approaches zero (Fig. 7). This causes a coherent specular forward reflection. Furthermore, most surfaces, including the sea surface, exhibit for $\alpha = 0°$, a reflection coefficient $\Gamma \approx -1$. The result of these two factors is that for a small grazing angle α, the forward reflected ray (path SPT) is almost equal in amplitude but out of phase with the direct ray (path ST). The two rays add destructively, causing a null in the target return when $\alpha = 0°$. More nulls occur whenever the difference between the direct path ST and the indirect path SPT is a multiple of the wavelength λ. These nulls correspond to a larger α. As α increases, the reflection becomes less specular, and the magnitude of the reflection coefficient becomes smaller than unit. The resulting nulls are therefore not as deep as the first null. Figure 8 presents a plot of a predicted radar power return (normalized with respect to the return at a distance of 1 km

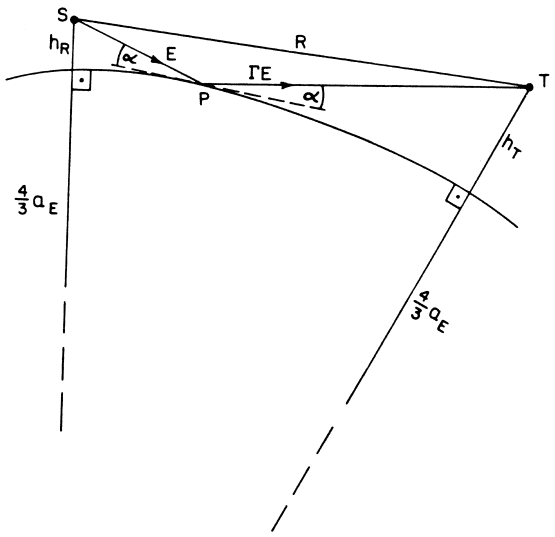


FIGURE 7 Multipath geometry.

in free space) as a function of range. The target and radar are both at a height of 100 m above the sea surface.

C. Atmospheric Attenuation

Propagation of radar signals at wavelengths shorter than 10 cm suffers attenuation due to the atmospheric constituents. Rain, fog, and dust particles attenuate the forward path by scattering the signal into other than the forward direction. The attenuation rate depends on the particle concentration. Water vapor and oxygen molecules attenuate by resonant absorption. Attenuation due to water vapor peaks at a wavelength of 1.35 cm, where the typical attenuation is 0.2 dB/km. Attenuation due to oxygen peaks at a wavelength of 0.5 cm, with a typical attenuation of 10 dB/km. At wavelengths longer than 10 cm, atmospheric attenuation is smaller than 0.01 dB/km and can be ignored.

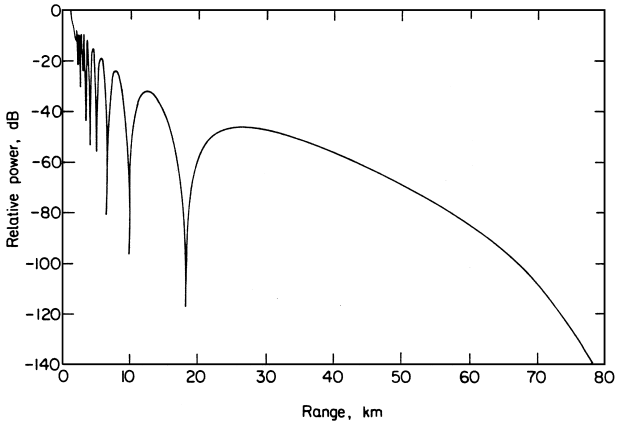


FIGURE 8 Received power as a function of range for a target and radar heights of 100 m above the sea surface.

D. Over-the-Horizon (OTH) Propagation

Radar instruments at wavelengths longer than 10 m utilize reflection from the ionosphere to reach ranges much beyond the horizon. In a similar way to short wave communication, OTH radars can reach ranges of thousands of kilometers. In OTH radars, the simple relation between delay and range (Eq. 1) does not apply, and corrections based on the ionosphere height are necessary.

VI. SURFACE BACKSCATTERING (CLUTTER)

Rough surfaces divert the reflected signal to other than the forward specular direction. Of special interest to radar is the backward reflection, better described as backscattering. Another name for surface backscattering is *clutter* because its effect is to clutter the radar display and musk target returns. Because clutter is reflected from extended surface area, it is customary to describe it in terms of the radar cross section per unit area of surface, known as the normalized radar cross section (NRCS). Its preferred symbol is σ°, and it is dimensionless.

σ° is strongly dependent on the surface roughness and the grazing angle α (also called depression angle). A perfectly rough surface should reflect back isotropically (but only to the upper hemisphere) and should exhibit $\sigma^\circ(\alpha) = 2$ for $0 < \alpha < 90^\circ$. Practical surfaces exhibit stronger backscattering when the backward direction coincides with the specular direction, that is, near $\alpha = 90^\circ$, decreasing with decreasing α. An empirical approximation of $\sigma^\circ(\alpha)$ for the sea surface at 4 m/sec wind and vertically polarized radar signal at frequency of 1 GHz is

$$\sigma^\circ(\alpha) = \begin{cases} 0.01\dfrac{\sin\alpha}{\cos^3\alpha}, & 0^\circ \le \alpha \le 85^\circ \\ \sigma^\circ(85^\circ), & 85^\circ \le \alpha \le 90^\circ \end{cases} \tag{24}$$

The radar reflection from the surface at any particular delay is obtained by integrating the backscattered signal from the entire area illuminated by the radar at that instant. Figure 9 presents a typical scenario of the area illuminated by a pulse radar at height h above the surface, with its antenna beam at a depression angle α. The across-range side of the illuminated area is determined by the horizontal beamwidth θ_H. However, the along-range side of the illuminated area is determined by the pulse width t_p. The effective target cross section of the illuminated area in Fig. 9 is

$$\sigma = \frac{C t_p R \theta_H}{2\cos\alpha}\sigma^\circ(\alpha). \tag{25}$$

Equation 25 assumes an ideal antenna of uniform gain within θ_H and zero gain outside the beam. In practical

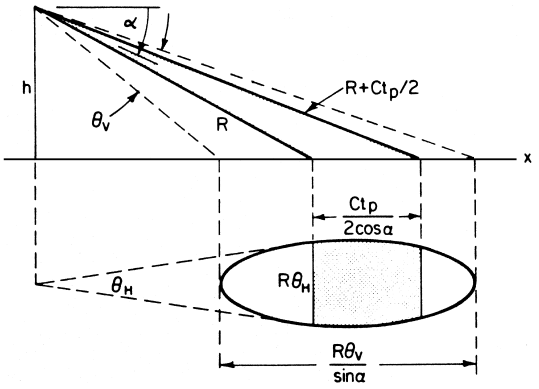


FIGURE 9 Pulse width limited ground illumination.

antennas, the gain drops gradually, and some residual gain (known as sidelobes) exists in all directions. Most harmful is that the direction in which the slant range to the ground is $R + CT_R/2$. From that patch of ground, an earlier pulse is returned. The problem of range ambiguity of $CT_R/2$ affects radars with small T_R, namely, with high pulse repetition frequency (PRF).

Clutter reduction is accomplished by minimizing the illuminated area. As Eq. 25 shows, this can be done by reducing the pulse width. Since the effective pulse width is the one measured in the receiver, after the matched filter, it is not necessary to reduce the actual duration of the pulse but the duration of the matched filter output. The matched filter output is the autocorrelation function of the transmitted pulse. Narrow autocorrelation is obtained from signals with a wide frequency spectrum. The design of such signals will be discussed in Section VII.

Further reduction of clutter can be achieved when either the radar or the target is in motion. In that case, a coherent radar can use frequency filtering to divide the reflection at a given delay (range) into smaller fractions according to their Doppler shift. With Doppler filtering, the radar resolution element gets an additional dimension—range rate.

VII. PULSE COMPRESSION

Good range resolution requires a narrow effective pulse. A high PD requires a large SNR. Optimizing both requirements, without increasing the transmitter peak power, is accomplished by transmitting an optimal signal for the application and utilizing a matched filter in the receiver.

A. Matched Filter

For a given radar signal $s(t)$ of finite duration t_p, maximum SNR at the receiver output is obtained if the receiver contains a matched filter whose impulse response $h(t)$ is the mirror image of the signal.

$$h(t) = Ks^*(t_p - t). \tag{26}$$

K is an arbitrary constant which takes care of dimensions. The asterisk (*) indicates that if the mathematical description of the signal is a complex function, then the mathematical description of the impulse response is the mirror image of the conjugate of the signal. The important property of the matched filter is to concentrate most of the energy of the signal at its end, at t_p. If the total signal energy is E, and if the noise is white with a power spectral density N_0, then at t_p the output SNR is the highest attainable SNR,

$$\mathrm{SNR}_{max} = \frac{2E}{N_0}. \tag{27}$$

Another important property of the matched filter is that the actual waveform of the output signal is identical to the autocorrelation function of the input signal. In order to get a short effective output signal, the input signal does not itself have to be short, but its autocorrelation function has to be effectively short. This can be accomplished by increasing the frequency bandwidth while maintaining the original duration. The ratio between the duration of the transmitted signal and the duration of the matched filter output signal is called the *compression ratio*, and the entire process is called pulse compression.

B. Linear-FM Signal

An example of a signal that can yield a large compression ratio is the linear-FM signal, shown in Fig. 10. The signal exhibits fixed amplitude during its duration t_p, but its instantaneous frequency changes linearly from $f(0)$ at $t = 0$ to $f(0) + \Delta f$ at $t = t_p$. The envelope of the matched filter output (the magnitude of the autocorrelation) for the case of $t_p \Delta f = 10$ is plotted in Fig. 11. The time axis is centered about the point of maximum output and is also normalized with respect to the pulse duration. The dashed line presents the autocorrelation of a fixed-frequency signal of

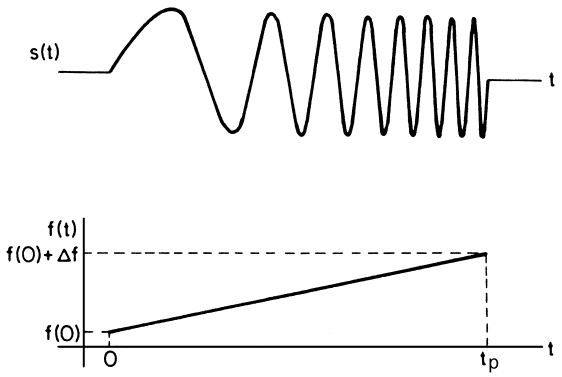


FIGURE 10 A linear-FM signal: (a) waveform, (b) frequency.

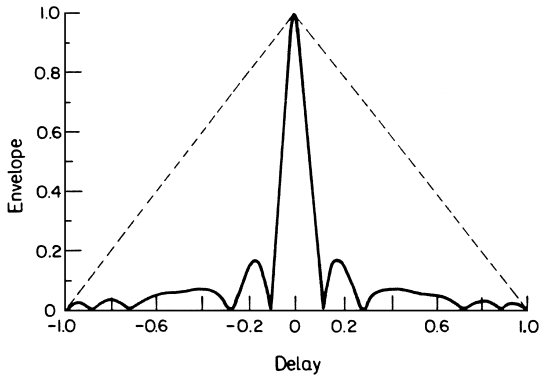


FIGURE 11 Matched filter response of a linear-FM signal.

the same duration t_p. Figure 11 demonstrates the improved range (delay) resolution of the linear-FM signal over the fixed-frequency signal. The width of the main lobe is reduced by the compression ratio $t_p \Delta f$. An ideal output signal should not have the sidelobes observed in the linear-FM response. The sidelobes can be partially reduced by either using more complicated modulation than the linear frequency modulation (e.g., amplitude and phase modulation) or by "unmatching" the receiving filter. Modifications of the matched filter must result in some loss in SNR. The sidelobe level can be reduced to any desired level by increasing the length of the filter's impulse response much beyond the duration of the input signal. Filters of this kind are called inverse filters.

C. Phase and Frequency Coding

Many pulse compression signals are based on splitting the transmitted pulse into M bits and phase or frequency modulating the bits according to a coded sequence. Amplitude modulation is rarely used because it cannot be easily implemented in high-power transmitters.

The simplest phase modulation switches between two phases, preferably 0° and 180°, which can be referred to as $+$ and $-$. Such a two-valued sequence is called a *binary sequence*. An important family of pulse compression binary sequences are the *Barker* codes. The largest Barker sequence is 13 bits long:

$$+ + + + + - - + + - + - +$$

The normalized envelope of its matched filter response is plotted in Fig. 12. It is characterized by sidelobes which do not exceed $1/M$. The pulse compression is obviously M.

Frequency coding can also be used to achieve pulse compression. The linear-FM modulation can be approximated by M discrete frequency steps. This code is called *quantized linear FM*. If the linear phase during each one of the M quantized frequencies is further quantized into M equally spaced phases, a polyphase code will result,

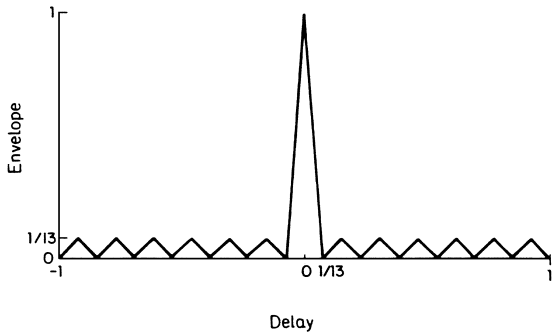


FIGURE 12 Matched filter response of the longest Barker signal.

which also has pulse compression properties. Known as *Frank* codes, these sequences can be very large and always M^2 bits long.

D. Frequency Resolution and the Ambiguity Function

When the received signal is Doppler shifted, the input to the matched filter differs from the signal to which it was matched. The response of the matched filter to its signal after Doppler shift is described by a two-dimensional extension of the autocorrelation function, known as the *ambiguity function*. An important property of the ambiguity function is that the volume underneath it is a constant. Among other things, it means that improved delay resolution must come at the expense of reduced Doppler resolution.

Range resolution was obtained by using a short pulse and by frequency or phase modulating it in order to increase its frequency bandwidth. On the other hand, velocity (Doppler) resolution can be improved by increasing the time duration of the pulse. One signal whose ambiguity function exhibits a narrow peak in both the delay and the Doppler axes is a *coherent train of M pulses*. With such a signal, the range resolution is proportional to the width of a single pulse t_p, while the Doppler resolution is inversely related to the duration of the entire train of pulses $1/(MT_R)$. The remaining volume of the ambiguity function reappears in many other peaks which pop up on a mesh, every T_R along the delay axis and every $1/T_R$ along the Doppler axis. An ambiguity function of this kind is known as "a bed of nails."

VIII. DOPPLER FILTERING

Doppler filtering is used to improve the radar resolution by adding the Doppler (velocity) dimension to the two other resolution dimensions—range and bearing. The most widely used signal for Doppler filtering is the coherent train of M pulses. Each pulse produces a sample

for every range cell. The M samples from a given range cell are used to determine the Doppler shift caused by the target at that range. The samples must be processed coherently, namely, contain both amplitude x and phase ϕ information. A preferred alternative to amplitude and phase are the in-phase I and quadrature Q components:

$$\begin{aligned} s(t) &= x(t)[\cos \omega_R t + \phi(t)] \\ &= I(t) \cos \omega_R t - Q(t) \sin \omega_R t, \; w_R \\ &= 2\pi f_R. \end{aligned} \tag{28}$$

The Doppler shift of the reflected signal at a given range cell can be deduced from the M pairs of I and Q by using discrete Fourier transform (DFT). Each output of the DFT represents a Doppler filter. The width of each filter is approximately equal to $1/(MT_R)$.

A. Moving Target Indicator (MTI)

There are applications in which not all the individual Doppler filters are needed, for example, stationary radars that try to resolve moving targets out of the relatively stationary ground clutter. Such an application employs a moving target indicator (MTI). The MTI can be performed using the DFT of M pulses by combining all the high-Doppler filters into one output. However, the MTI can be performed with as little as two pulses. A simple MTI, the *single canceller*, is shown in Fig. 13a. It must follow a phase detector rather than an envelope detector. The circuit subtracts the echo of the previous pulse from the echo of the present pulse. The existence of Doppler shift will result in a phase difference between the two echoes, and subtraction will not cancel them out. The Doppler frequency response of the single canceller is plotted (solid line) in Fig. 13b. The null at zero-Doppler reflects the clutter cancellation property. The null may be too narrow for situations in which the ground clutter undergoes some motion (e.g., sea waves). The null can be broadened by cascading two such stages to create the *double canceller*. The clutter attenuation caused by the MTI circuitry is characterized as *sub-clutter visibility* (SCV), or *improvement factor*.

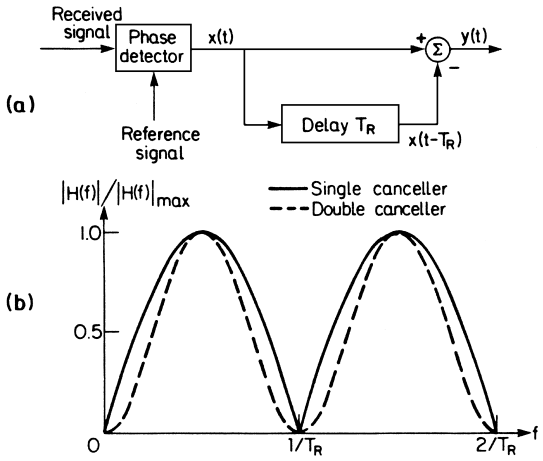


FIGURE 13 The MTI: (a) single canceller, (b) frequency response.

As seen in Fig. 13b, MTI cancelers also attenuate Doppler frequencies which are multiples of the PRF. The corresponding targets' radial velocities are known as *blind speeds*, because targets at these velocities cannot be detected. This problem can be resolved by changing the pulse interval T_R, creating what is known as *staggered PRF*.

B. Synthetic Aperature Radar (SAR)

Synthetic aperature radar (SAR) is a more elaborate extension of Doppler filtering. It can be implemented in airborne radars to yield improved across-range resolution. When a radar is flown on a straight course, its distance from a point on the ground is a quadratic function of time (parabola). Minimum range occurs at the time of closest approach, when the range vector (to the point on the ground) is perpendicular to the flight course. The phase of a coherent radar signal reflected from that point is also a quadratic function of time, because the phase is linearly related to range

$$\phi(R) = 2R\frac{2\pi}{\lambda} + \phi_0. \tag{29}$$

The idea behind SAR processing is to compensate for the expected phase history and thus cause constructive addition only of returns from that particular point on the ground. For comparison, in a fixed-frequency filter like a DFT, the compensating phase is a linear function of time.

In a side-looking SAR, the improved resolution due to this kind of processing is in the across-range direction, while the radial resolution is obtained using pulse compression. The across-range resolution element is

$$\Delta y = \frac{\lambda}{2L} R, \tag{30}$$

where L is the distance traveled by the airborne radar while coherent processing is taking place. Recall from Fig. 9 that in a conventional radar the across-range resolution is $\theta_H R$. Comparing resolutions shows that the effective horizontal beamwidth in SAR is $\lambda/2L$, or that the effective antenna aperture is proportional to L. This property is what gave SAR its name. L is limited by the requirement that throughout L the ground pixel must remain within the actual antenna beam. SAR is usually used to produce a high-resolution map of a wide ground swath parallel to the flight path. Thus, the signals from many pixels must be

processed simultaneously. A large L further increases the amount of parallel processing. In applications like spaceborne SAR, the processing requirements may be too large for real-time analysis.

IX. ANTENNAS AND TRACKING

The radar antenna is the key to direction measurement and angular resolution. Large aperture antennas provide both directivity (narrow beam) and high gain—two properties that are important in most radar applications. Large aperture by itself cannot produce a narrow beam. The phase of the radiation at every point in the plane of the aperture must be the same. There are two basic approaches to achieve equal phase on a large plane. The first approach utilizes a single small radiator, known as the feed, which illuminates a large reflector in such a way that the reflected radiation will have the property of equal phase along a large plane. The most common implementation is a *parabolic reflector*, with a feed at its focus (Fig. 14a). It is well known that the distance from focus to parabola to plane across the parabola is the same for every path possible. Parabolic reflectors are simple to construct. Polarization and frequency bandwidth are determined by the feed, which can be easily changed. Steering the beam must be mechanical, although for small shifts it is sufficient to move only the feed. The second approach is the *planar array* antenna—an array of small radiators arranged on a plane. The radiators are usually dipole antennas in front of a plane reflector or slots in the front wall of a cavity. The equiphase front of a planar array is parallel to the array plane, when all the radiators are fed with equal-phase signals. The beam can be steered (Fig. 14b) to an angle θ by shifting the phase to the nth element of an array of equal spacing s,

$$\phi_n = ns\frac{2\pi}{\lambda}\sin\theta. \tag{31}$$

Phase shifting can be performed electronically at a very high speed. This permits fast electronic beam steering. Because of the relation between direction and phase, such arrays are also called *phased arrays*. This name may be more appropriate than planner arrays because modern phased arrays are not necessarily mounted on a plane. The array can conform to the curved surface of its carrier.

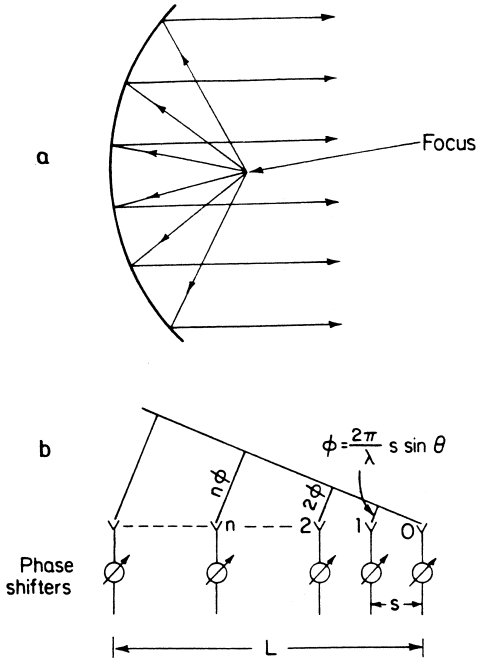


FIGURE 14 Radar antennas: (a) parabolic reflector, (b) phased array.

A. Main Beam and Sidelobes

The angular dependence of the far field of a one-dimensional array of length L is

$$E(\theta) = E(0)\frac{\sin\left(\frac{\pi L}{\lambda}\sin\theta\right)}{\frac{\pi L}{\lambda}\sin\theta}. \tag{32}$$

This is the well-known "sine over argument" pattern which exhibits a main lobe of "half power" width

$$\theta_{3\,\mathrm{dB}} = 0.886\frac{\lambda}{L}\ \text{(radians)}. \tag{33}$$

The omnipresent sidelobes decay with increasing θ but never completely disappear. The first and highest sidelobe is only 13.5 dB lower than the main lobe. Sidelobes can cause severe problems in radar, and their reduction or blanking are of paramount importance. Feeding the array elements with nonuniform intensity can reduce the sidelobe level.

B. Antenna Tracking

Radar antennas operate in two modes: scannig and tracking. In the scanning mode, all the targets in a section are illuminated periodically. In the tracking mode, the antenna is dedicated to one target and automatically follows its changing direction. Automatic tracking requires an ability to measure the angle error and a feedback loop that will use this error information to correct the antenna direction. The angle error is the angle between the peak of the antenna beam and the direction to the target. In high SNR it can be determined to within a very small fraction of the beamwidth. There are several techniques to accomplish that. All of them are based on effectively performing a derivative operation on the antenna pattern. A derivative

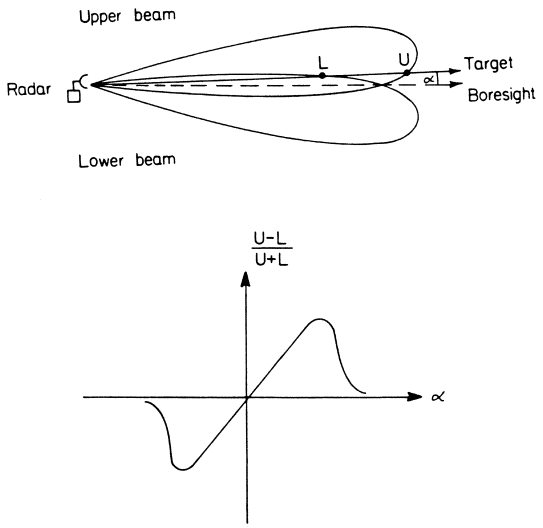


FIGURE 15 Monopulse antenna tracking.

near a maximum is relative to the distance from the maximum (including the sign) and can serve as an error signal. The derivative of a function is obtained by subtracting two slightly shifted copies of the function. The various tracking methods differ in how they implement that subtraction. The most important is the *monopulse* technique. Two beams in opposite offsets from boresight are created using two off-axis feeds. The received signals from the two beams are subtracted, creating a "difference" signal. The difference signal, normalized with respect to a "sum" signal, is proportional to the angular error (Fig. 15). The two beams must be well balanced, so that the difference signal disappears exactly when the target is on the boresight line of the antenna. The name monopulse stems from the fact that the two offset beams are available simultaneously, due to the use of two feeds and two receiving channels. Thus, angle measurement can be performed using a single reflected pulse. In a different approach, the two offset beams are created sequentially by alternating the direction of the beam every few pulses. This method is called *sequential lobing*. Note that in both methods two offset beams are required for each of the two axes (azimuth and elevation). In a third method, called *conical scan*, a single offset beam is rotated around the boresight direction, providing angle error information in both axes.

X. SPECIAL RADARS

A. Continuous Wave Radar

A continuous wave (CW) radar, as its name implies, emits a continuous signal. It must therefore receive the returned signal while transmitting. The unavoidable leakage between transmitter and receiver means that the weak reflected signal may have to compete with the strong directly received transmission. Separation between the two must be based on parameters other than intensity.

The CW design is found in radars that emphasize velocity measurement, such as police radars or artillery muzzle velocity radars. The Doppler shift provides the means to separate the transmitted signal from the received signal. The long signal duration enables high-resolution velocity measurement. For a CW radar to be able to measure range too, the transmitted signal must be marked on the time axis. Such marking is usually implemented through periodic phase or frequency modulation. The modulation also helps to separate the target-reflected signal from the directly received signal. Still, it is important to minimize the direct reception, which is why CW radars usually use two separate antennas, a transmitting one and a receiving one.

A major advantage of CW radars is pointed out in Eq. 7, which shows that the SNR is a function of the average transmitted power during target illumination. A pulse radar needs high peak power to achieve sufficient average power, while in a CW radar the peak power is equal to the average power. This is why CW radars use low-power transmitters, based on low-voltage solid state devices rather than on high-voltage vacuum tubes.

B. Secondary Surveillance Radar

It is in the interest of targets to enhance their response to friendly radars. Passively, this can be done by using retroflectors. A stronger effect is reached when the target employs an active transponder which amplifies and retransmits the radar signal that hits it. A retransmitted signal can be much stronger than a reflected signal. It can also be modified before retransmitting. A frequency shift will separate it from clutter. In air traffic radars, the transponder shifts the carrier frequency 60 MHz up and modulates it with information about the airplane identity and barometric altitude.

C. Bistatic Radar

The radar receiver is not always collocated with the transmitter. There are situations in which it is advantageous to use an illuminating transmitter in one location and one or more receivers in other locations. An important application is semiactive missile homing. The missile carries a receiver and an antenna. The missile-borne receiver gets two receptions of the same radar transmission, one through the radar-missile path and one through the radar-target-missile path. Delay and Doppler differences between the two, and the angle to the target, as well as their rate of change, can all, or partly, be used for homing purposes. In addition to simplifying the missile-borne payload, the

bistatic approach allows the missile to remain electromagnetically passive and therefore to be less trackable.

Target reflection characteristics are different in bistatic situations. The reflection is somewhere between backward and forward scattering. A target smoothed to minimize backward reflection may yield stronger bistatic reflection at some favorable combination of angles.

SEE ALSO THE FOLLOWING ARTICLES

ELECTROMAGNETICS • PLANETARY RADAR ASTRONOMY • REMOTE SENSING FROM SATELLITES • SYNTHETIC APERTURE RADAR

BIBLIOGRAPHY

Bowen, E. G. (1987). "Radar Days," Adam Hilger, Bristol, UK.
Barton, D., and Leonov, S. (1997). "Radar Technology Encyclopedia," Artech House, Norwood, MA.
Curry, G. R. (2001). "Radar System Performance Modeling," Artech House, Norwood, MA.
Franceschetti, G. (1999). "Synthetic Aperature Radar Processing," CRC Press, Boca Raton, FL.
Henderson, F. M., ed. (1998). "Manual of Remote Sensing, Vol. 2, Principles and Applications of Imaging Radar," Wiley, New York.
Kawecki, J. M., ed. (1999). "Radar Essentials: Selected Readings," IEEE Press, New York.
Levanon, N. (1988). "Radar Principles," Wiley, New York.
Nathanson, R. (1999). "Radar Design Principles: Signal Processing and the Environment," 2nd ed., SciTech, Mendham, New Jersey.
Peebles, P. Z. (1998). "Radar Principles," Wiley, New York.
Skolnik, M. L., ed. (1990). "Radar Handbook," 2nd ed., McGraw-Hill, New York.
Stimson, G. W. (1983). "Introduction to Airborne Radar," Hughes Aircraft Co., El Segundo, CA.
Skolnik, M. (2001). "Introduction to Radar Systems," 3rd ed., McGraw-Hill, New York.
Son, T. (2001). "Range-Doppler Radar Imaging and Motion Compensation," Artech House, Norwood, MA.
Taylor, J. D. (2001). "Ultra-Wideband Radar Technology," CRC Press, Boca Raton, FL.



Radio Propagation

Henry L. Bertoni
Polytechnic University

I. Introduction
II. Influence of Terrain on Propagation
III. Atmospheric Effects
IV. Influence of Buildings
V. Influence of Vegetation

GLOSSARY

Antenna directivity The variation with direction of the power density radiated by an antenna, relative to isotropic radiation of the same total power.

Brewster's angle The angle of incidence of a plane wave in air for which there is total transmission into a dielectric for polarization of the magnetic field parallel to the surface of the dielectric.

Delay spread The spreading in time of received pulses due to echo. Usually measured by the second moment of the time dependent received power.

Diffraction Wave process, other than reflection and refraction, by which geometrical rays, and their associated fields, are scattered into many directions by an obstacle.

Fast fading Variations of the received power over distances on the order of a wavelength as a result of multipath. Also referred to a Rayleigh fading.

Fresnel zone Volume about a ray path in which the fields propagate. For a direct ray is an ellipsoid of revolution.

Ionosphere The part of the atmosphere above about 60 km where there are free electrons due to ionization by solar radiation.

Multipath The condition of having the electromagnetic waves arriving at a point in space along many different ray paths.

Path gain Ratio of the power received when the receiving antenna is match terminated to the transmitted power in the absence of antenna losses. The path loss is the negative of the path gain expressed in deci-Bells (dB).

Polarizability The property of a dielectric that contributes to the electric displacement. One plus the polarizability is the relative dielectric constant.

Range index The exponent of the distance between antennas that gives the variation with distance of the overall average received power. In free space the range index is 2.

Rays The paths in space used to localize the region through which electromagnetic fields reach a point when the wavelength is much smaller than the path length.

Shadow fading Spatial variations of the received power over a scale on the order of 10 m as a result of shadowing by buildings, trees, terrain, etc. Also known as Lognormal fading or slow fading.

Troposphere The part of the atmosphere below about 15 km where temperature normally decreases with altitude.

WIRELESS COMMUNICATION was originally developed by Marconi and others at the turn of the 20th century for long-distance (100s to 1000s of km) communication between fixed terminals and to ships. One hundred years later, a dense network of fiber optic and wired links span the world, providing much of the communication between fixed terminals. During the last 50 years, advances in microelectronics reduced the size and power consumption of radios, permitted use of higher frequencies to increase the rate of information transmission, and allowed implementation of complex schemes for signal coding and network control. These technological advances, together with our thirst for communications, produced the widely popular cellular mobile telephone network. Mobile phones, and other existing and proposed mobile systems, use radio links to cover relatively short distances (1 m to 10s of km) and rely on the dense fixed network to provide the rest of the communication link. Thus the applications with the greatest commercial significance evolved from low-frequency communication over long distances to high-frequency communication over short distances. However, high-frequency satellite links will continue to be important for communication when the distances are necessarily large, as in the case of communications to ships at sea or airplanes, and for applications such as the Global Positioning System (GPS) and Digital Audio Broadcast. Also, point-to-point microwave links will continue to provide cost-effective communications in many applications.

I. INTRODUCTION

Radio waves of frequencies below 3 MHz have wavelength that is greater than 100 m, and hence large compared to the height of most buildings and small hills. Propagation of these waves over terrestrial links is influenced by the earth's ionosphere and by the electrical properties and curvature of the earth (Wait, 1996), but only weakly by buildings and vegetation (Hall and Barclay, 1989). For frequencies above 30 MHz, the wavelength is less than 10 m, and hence smaller than buildings, trees, and terrain undulation. Propagation of these high-frequency waves over short distances is therefore strongly influenced by the buildings, trees, and terrain in the vicinity of, and between, the transmitter and receiver. While the ionosphere is not important for high frequencies, the variation of refractive index in the atmosphere is important over distances greater than about 10 km. Radio waves having frequency in the middle range between 3 and 30 MHz may be influenced by all features along the path.

In radio communications, a transmitting antenna is used to launch an electromagnetic signal into space, which is received by a second antenna at a remote location. For antennas located in free space, such as for communication between satellites, the propagation effects consist simply of the reduction of the signal with distance as a result of spreading the transmitted energy over an area that increases with distance. When one or both ends of the communication link are located near the ground, the atmosphere, terrain variation, vegetation, and buildings influence the received signal. Besides the dependence on the separation between transmitter and receiver, these features effect signal characteristics such as spatial and temporal fading, frequency dependence, and echo-induced time delay spread. This chapter describes how the individual features influence the characteristics of the received signal. Ultimately, it will be necessary to develop a description of how these features interact to simultaneously influence the radio wave.

II. INFLUENCE OF TERRAIN ON PROPAGATION

In order to describe the influence of terrain, and other features, along the propagation path, we use as a reference signal properties for antennas in free space. Antennas used for low frequencies (below 3 MHz) are large and therefore mounted near the earth, which has a strong influence on their radiation properties. Such antennas excite ground waves that can propagate around the curvature of the earth (Hall and Barclay, 1989; Wait, 1996), but are not strongly influenced by terrain irregularity. At high frequencies (above 30 MHz) the antennas are smaller and can be mounted above ground, so that they initially radiate as if in free space. The presence of the ground modifies the radiated fields through the processes of reflection and diffraction at terrain features, such as hills. This section will focus on the propagation of high frequency waves.

A. Path Gain/Loss for Antennas in Free Space

When the region around an antenna is free of obstacles, the electromagnetic fields radiated by the antenna propagate in the form of a spherical wave. The far field region comprises radial distances r from the antenna that are larger than about $2L^2/\lambda$, where L is the largest dimension of the antenna and λ is the wavelength (Balanis, 1997). In this region, the radiated electric field is given by (Bertoni, 2000)

$$\mathbf{E} = \mathbf{a}_r \sqrt{\frac{\eta P_T}{2\pi}} \frac{e^{-jkr}}{r} \sqrt{g(\theta, \phi)}. \tag{1}$$

Here $\mathbf{a}_r$ is a unit vector describing the polarization of the electric field, which is perpendicular to the direction of propagation, and phase not included in the exponential factor. In this expression $k = 2\pi/\lambda$ is the wavenumber,

$\eta \approx 120\pi$ ohms is the wave impedance and P_T is the total radiated power. The directivity $g(\theta, \phi)$ describes the variation of the fields with the direction of propagation (θ, ϕ) shown in Fig. 1, and is normalized such that its integral over the unit sphere is 4π.

The magnetic field $\mathbf{H}$ is perpendicular to $\mathbf{E}$ and to the radial so that the Poynting vector $\mathbf{P} = \frac{1}{2}\mathrm{Re}\{\mathbf{E} \times \mathbf{H}^*\}$, which represents the power density crossing the spherical surface at the distance r, is in the outward radial direction. The power density, which is given by $\mathbf{P} = \mathbf{a}_r P_T g(\theta, \phi)/(4\pi r^2)$, is the total radiated power divided by the area of the sphere, with the directivity giving the deviation of the power from the average. Neglecting dissipation in the antenna, the maximum value of $g(\theta, \phi)$ is the antenna gain G. Because the electric and magnetic fields are transverse to the direction of propagation, symmetry precludes there being an antenna that is an isotropic radiator. However, if such antenna could exist, then $g(\theta, \phi)$ would be unity, the power would be uniformly distributed over the sphere, and $G = 1$. Although no real antenna can be isotropic, we will make use of this convenient concept.

When receiving signals, an antenna has an effective capture area A_e such that the received power is the product of A_e and the incident power density when the antenna terminals are connected to a matched load. It can be shown that for any antenna the effective area is related to the directivity $g(\theta, \phi)$ in the direction of the arriving wave through the formula $A_e = (\lambda^2/4\pi)g$. Isotropic antennas, for which $g = 1$, have an effective area $A_e = 0.08\lambda^2$. For communication between two antennas having the same polarization of the electric and magnetic fields, as shown in Fig. 1, the received power is given by $P_R = |\mathbf{P}|A_e$.

It is conventional to define the path gain PG as the ratio of the received power P_R to the transmitted power P_T. Since $P_R < P_T$, the path gain $PG = P_R/P_T$ is less than unity. The path gain for antennas in free space is

$$PG \equiv P_R/P_T = g_1 g_2(\lambda/4\pi r)^2, \qquad (2)$$

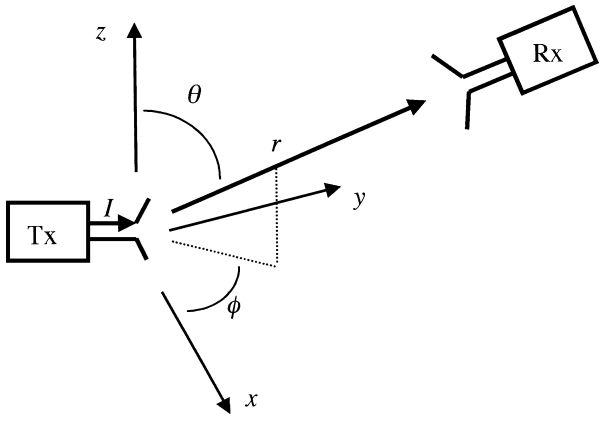


FIGURE 1 Transmitting and receiving antennas in free space.

where g_1, g_2 are the directivity of the two antennas in the directions of the connecting line. The path gain decreases with distance squared, and because it is proportional to λ^2, it also decreases with frequency as $1/f^2$. It is seen from (2) that the path gain is reciprocal in that the same value is obtained no matter which antenna is the transmitter. The path loss L in deci-Bell (dB) units is found using the expression $L = -10\log(\mathrm{PG})$, and increases with the separation between the antennas. When the antennas are located in the presence of the atmosphere, terrain, vegetation, or buildings, the path gain can be significantly altered from the value given by Eq. (2) for antennas in free space.

B. Antennas Over Flat Earth

For terrestrial links, where the antennas are placed above the earth, the earth acts as a reflector of the fields. For short links we may neglect the earth's curvature, so that the geometry is as shown in Fig. 2 for antennas separated by a horizontal distance R. If the antenna on the left is taken to be the transmitter, then it sends out a spherical wave that is reflected from the ground, as well as propagating directly to the receiving antenna. We can think of the spherical wave as traveling along a family of radial lines, called rays, emanating from the transmitter. The power density (2) of the spherical wave implies that energy is conserved in any cone (tube) of rays. Moreover, in a small volume about a ray the electric and magnetic fields of the spherical wave (1) are like those of a plane wave. As a result the ray fields that are incident on the earth will approximately be reflected according to Snell's law, and their amplitude will be reduced by the same reflection coefficient Γ found for plane waves (Jordan and Balmain, 1968; Wait, 1996). Because the angle of reflection of the ray will be the same as the angle of incidence, the rays reflected from a flat surface will appear to come from the image of the transmitter in the ground plane.

We may use (2-1) for the direct ray, and for the reflected ray by including Γ. For simplicity we neglect the differences in the directivity of each antenna for the direct and

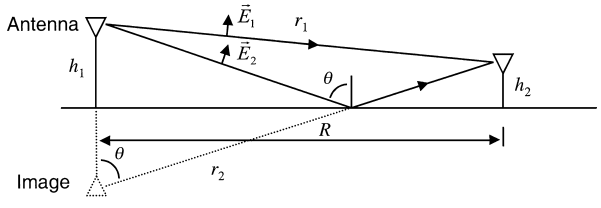


FIGURE 2 Direct and reflected ray paths for antennas located above a plane earth.

reflected rays. Then the path gain (2) resulting from the sum of the fields will be

$$PG = g_1 g_2 \left(\frac{\lambda}{4\pi}\right)^2 \left| \frac{e^{-jkr_1}}{r_1} + \Gamma(\theta_2)\frac{e^{-jkr_2}}{r_2} \right|^2. \quad (3)$$

Here the path lengths r_1 and r_2 are given by the formula $r_{1,2} = \sqrt{R^2 + (h_1 \mp h_2)^2}$. As noted above, when the antenna is close to the ground in terms of wavelengths, it may be necessary to account for an additional ground wave contribution (Wait, 1996).

The path gain obtained from (3) for isotropic antennas ($g_1 = g_2 = 1$) is plotted on a dB scale in Fig. 3 using the reflection coefficients Γ_E for horizontal (TE) polarization, Γ_H vertical (TM) polarization, and $\Gamma = -1$ for a perfect conductor. The plots are for base station height $h_1 = 8.7$ m, mobile height $h_2 = 1.8$ m, and frequency $f = 900$ MHz. For R ranging from 2 to about 150 m in Fig. 3, the path gain varies due to the interference between the two terms in (3) as they go in and out of phase. As R approaches 200 m, the path distances r_1 and r_2 approach each other and the interference minima get deeper for $\Gamma = -1$. For horizontal (TE) polarization, Γ_E is negative for all angles of incidence, and its magnitude increases monotonically to unity with θ. As a result, the path gain in Fig. 3 for horizontal polarization is similar to that for $\Gamma = -1$. For vertical (TM) polarization, Γ_H has a zero at the Brewster angle θ_B, which occurs at the point shown in Fig. 3. For smaller values of R, $\theta < \theta_B$ and Γ_H is positive, with the consequence that the minima for vertical polarization occur at the maxima for the case $\Gamma = -1$. Beyond the Brewster angle point, Γ_H is negative and the interference maxima and minima for vertical polarization occur at the same locations as for the horizontal polarization, although they are not as deep since $|\Gamma_H|$ is smaller than $|\Gamma_E|$.

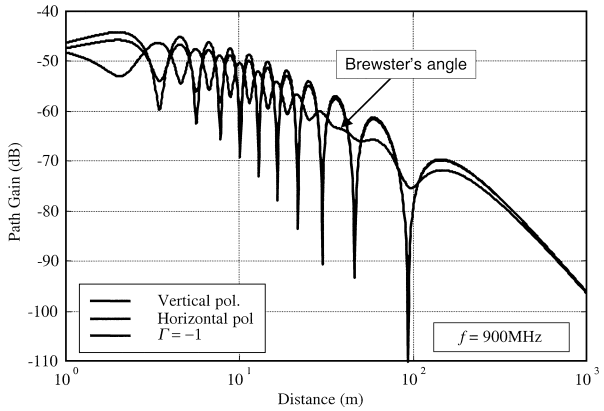


FIGURE 3 Path gain computed from the two ray model using the plane wave reflection coefficients for horizontal (TE) and vertical (TM) reflection coefficients, and using $\Gamma = -1$ ($h_1 = 8.7$ m, $h_2 = 1.8$ m, $f = 900$ MHz). (From Bertoni, H. L. (1999).)

For typical terrestrial applications, $R \gg h_{1,2}$, so that θ approaches 90° and $\Gamma \approx -1$. Using the quadratic approximation for $r_{1,2}$ we obtain

$$PG \approx 4\left(\frac{\lambda}{4\pi R}\right)^2 \left| \sin\left(2\pi \frac{h_1 h_2}{\lambda R}\right) \right|^2. \quad (4)$$

It is seen from (4) that raising the height of either antenna for a fixed R results in a sinusoidal variation of the received power. As R increases, the argument of the sine function in (4) decreases, causing the maxima and minima observed in Fig. 3, until the argument reaches $\pi/2$ and the last maximum is obtained. This last maximum occurs at the break point R_b given by $R_b = 4h_1 h_2/\lambda$. Beyond the break point, the argument of the sine function in (4) becomes small and the sine can be approximated by its argument. Thus for $R \gg R_b$, the path gain reduces to $PG \approx (h_1 h_2)^2/R^4$. From this formula, it is seen that doubling the height of either antenna increases the received signal by a factor of 4, or 6 dB. Beyond the break point the presence of a flat earth causes the received signal to vary as $1/R^4$, instead of the $1/R^2$ as found in free space, and the path gain to be independent of frequency. In later sections it will be found that the $1/R^4$ variation generally occurs whenever the propagation path is nearly horizontal.

C. Knife Edge Diffraction and Fresnel Zones

The basic diffraction mechanism involved in propagation over terrain obstacles is most easily understood in terms of propagation past an absorbing knife edge, which is depicted in Fig. 4. Treating the antenna on the left as the transmitter, it radiates rays whose fields $\mathbf{E}_{GO}$ illuminate the knife edge. One of the rays is incident on the edge itself, where it generates diffracted ray fields $\mathbf{E}_D$ that appear to come from the edge. Continuing this ray to the right defines the shadow boundary of the incident fields. In the illuminated region above the shadow boundary, both the

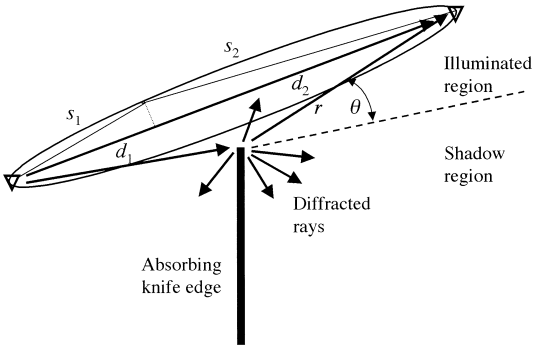


FIGURE 4 Diffracted rays generated at a knife edge, and the Fresnel zone about the direct ray from the transmitter to a receiver in the illuminated region.

direct rays from the transmitter and the diffracted rays contribute to the total field, while in the shadow region below the shadow boundary only the diffracted wave contribute.

For normal incidence on the edge, the diffracted rays are generated at the edge and leave in all directions in the plane perpendicular to the edge (called the plane of incidence). Near the edge, the diffracted fields take the form of a cylindrical wave. The diffracted rays also spread in the direction perpendicular to the plane of Fig. 4, with the transmitting antenna acting as the common point of spreading. As a result, the field diffracted by an absorbing edge is given by (Felsen and Marcuvitz, 1973; McNamara *et al.*, 1990)

$$\mathbf{E}_D = \mathbf{E}_{GO}(0)\frac{\exp(-jkr_2)}{\sqrt{r_2}}\sqrt{\frac{r_1}{r_1+r_2}}D(\theta). \qquad (5)$$

Here $E_{GO}(0)$ is the field incident on the edge and r_1, r_2 are the distances shown in Fig. 4, and $D(\theta)$ is the diffraction coefficient. Away from the shadow boundary, the magnitude of the diffraction coefficient for an absorbing knife edge (Felsen and Marcuvitz, 1973) is given by $|D(\theta)| \approx \sqrt{\lambda}/(2\pi|\theta|)$ for θ in radians. Because of the dependence of D on λ, the received power in the shadow region will decrease more rapidly with frequency than for the same antennas in free space. As a result of the dependence of D on θ, the receive power in the shadow region decreases rapidly as the receiver moves into the shadow region.

In the transition region about the shadow boundary, the distinction between the direct and diffracted fields becomes blurred (McNamara *et al.*, 1990). In Fig. 5 we have plotted the variation of the total field, normalized to the field incident on the edge, for a transmitter at a large distance. The field is plotted as a function of the perpendicular distance y above or below the shadow boundary at a distance of 30 m beyond the edge. Above the shadow boundary ($y > 0$), the diffracted fields interfere with the direct fields causing a small oscillation about the value that would occur in the absence of the edge. Right on the shadow boundary ($y = 0$), the total field is one half the fields in the absence of the knife edge, so that the received power is one quarter, or −6 dB. Below the shadow boundary only the diffracted field contributes, and its amplitude decreases with increasing angle away from the shadow boundary.

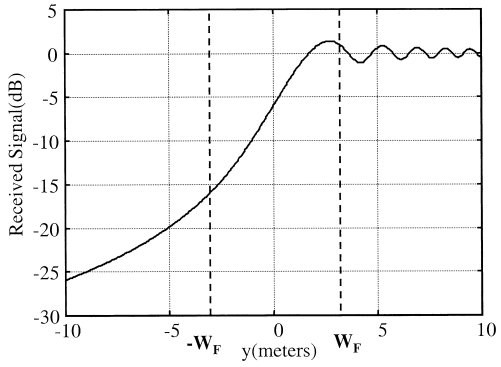


FIGURE 5 Variation of the total field, normalized to the field incident on the edge, about the shadow boundary 30 m from an absorbing knife edge due to a distant transmitter operating at 900 MHz. The plot is versus the perpendicular distance y from the shadow boundary and W_F is the width of the Fresnel zone. (From Bertoni, H. L. (1999).)

The Fresnel ellipse about the direct line from source to receiver determines the width of the transition region about the shadow boundary. Consider an ellipsoid of revolution defined by the condition that the distance $s_1 + s_2$ shown in Fig. 4 from the transmitter to any point on the surface and back to the receiver is $\lambda/2$ greater than the direct distance between the antennas. The Fresnel zone is the volume within the ellipsoid, whose cross section is shown in Fig. 4. If d_1 and d_2 are the distances from the antennas to any point along the line between them, then the half width W_F of the Fresnel zone at that point is $W_F = \sqrt{\lambda d_1 d_2/(d_1 + d_2)}$. In the case of Fig. 5, $d_1 \gg d_2$ so that $W_F = \sqrt{\lambda d_2}$. For $d_2 = 30$ m and $\lambda = 1/3$ m, $W_F = 3.16$ m, and the transition region in Fig. 5 extends ±3.16 m on either side of the shadow boundary.

The fields arriving at a point in space do so by propagating primarily within the Fresnel zone. If the Fresnel zone is unblocked ($y > W_F$), as shown in Fig. 4, the presence of the edge generates diffracted fields that adds to the incident field. When the edge is within the Fresnel zone ($-W_F < y < W_F$), the diffracted fields cannot be separated from the incident fields in a simple way and the UTD diffraction coefficient must be used (McNamara *et al.*, 1990). When the Fresnel zone is completely blocked, ($y < W_F$), the direct fields are prevented from reaching the point and only the diffracted fields contribute to the received power. At the middle of a ray path of length r, where the Fresnel zone is widest, the half-width is $W_F = \sqrt{\lambda r}/2$. As an example, at a frequency of 1 GHz and path length of 10 km, the maximum half-width is 27.4 m.

D. Diffraction Over Terrain Obstacles

Terrain irregularities range from small variations, such as hills, to mountains. When predicting path gain, the terrain variations are divided into large-scale features that are treated deterministically, and small-scale features whose effects are viewed statistically, if at all. For example, as part of an extensive measurement program Okumura *et al.* (1968) made measurements on rolling terrain. From the measurements they extracted the reduction of the average received power, and the increase in its variability, as

a function of the terrain roughness. Deterministic computations of terrain effects use the terrain profile along the propagation path. The ray optical theory diffraction is one approach to account for terrain features (Bullington, 1977; Luebbers, 1984). Full wave approaches, such as the parabolic equation method (Janaswamy and Andersen, 1998) can also be used, but are more numerically intensive.

The case of an isolated hill having wedge shape is suggested in Fig. 6. The antenna on the left transmits waves that illuminate the region above the wedge, including the edge. The edge then acts as a source of diffracted rays. For high frequencies, the Fresnel zone is narrow compared to the horizontal extent of terrain features, so that terrain variation transverse to the propagation path does not strongly affect propagation. Rays reflected from the left face of the wedge have their own shadow boundary that lies above the incident wave shadow boundary. Because the diffracted fields give only a small additional contribution to the fields in the illuminated region, they can be neglected in computing the path loss to receivers located there. When the receiving antenna is below the shadow boundary, the path gain is

$$PG = g_1 g_2 \left(\frac{\lambda}{4\pi}\right)^2 \left[\frac{|D|}{\sqrt{r_1 r_2 (r_1 + r_2)}}\right]^2. \qquad (6)$$

The diffraction coefficient D in (6) depends on the boundary conditions, polarization, wedge angle, and the angles ϕ and ϕ' in Fig. 6. In the transition region, the dependence of the received power is as discussed earlier for diffraction by a knife edge. Outside the transition region, but not near the shadow boundary of the field reflected from the face of the wedge, the dependence of D is similar to that for a knife edge.

In addition to the ray paths that are diffracted directly over an edge, there can be paths that involve ground reflection as well as diffraction. Figure 6 shows the images of the two antennas in the ground that can be used to account for reflections near the transmitter and the receiver. The ray fields that undergo reflection from the earth at either or both ends of the radio link add to the fields going directly from the transmitter to the receiver to determine the total received power. When summing the fields it is necessary to account for the phase differences for the various ray paths.

Multiple hills may intercept the path, longer propagation paths, diffracting the fields from hilltop to hilltop. In this case fields can be found from a cascade of the diffraction expressions at each hill. Various simplified treatments of multiple diffraction over hills have been developed to assist in numerical computation (Du Castel, 1966). For diffraction over rounded hills it may be necessary to account for creeping rays (Felsen and Marcuvitz, 1973) when finding the equivalent diffraction coefficient. Diffraction can describe propagation around the curvature of the earth (Wait, 1996). At high frequencies, the diffraction loss is substantially higher than the measured loss. Instead, the propagation of the horizon is due to atmospheric effects. Very low-frequency waves have much smaller diffraction loss, but the earth's ionosphere is equally important for trans-horizon propagation (Wait, 1996). The various phenomena associated with the earth's atmosphere that influence long-distance propagation are discussed in the next section.

III. ATMOSPHERIC EFFECTS

The earth's atmosphere has various effects on radio propagation. Absorption of the radio wave energy by oxygen, water vapor and other gases, and by rain, reduces the received signal at very high frequencies. Change in the atmospheric density with height produce a height variation of the refractive index, which bends rays propagating near to horizontal. Ray bending can significantly affect the received power beyond the horizon at high frequencies (above 30 MHz). Gravity produces a standard background variation of refractive index that tends to bend the rays around the curvature of the earth, therefore enhancing the received power. Variations in temperature and water vapor content with altitude create additional variations that can either enhance or reduce the received power depending on conditions. Scattering by small-scale variations in the refractive index due to turbulence in the troposphere

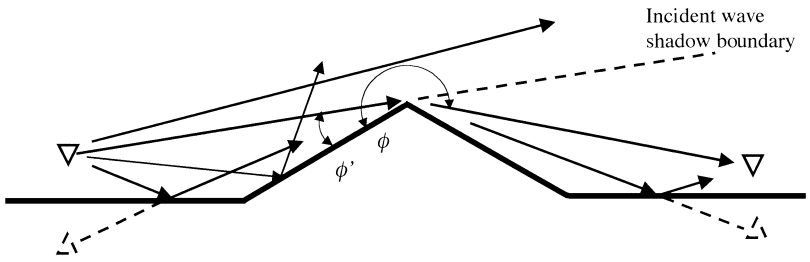


FIGURE 6 Diffraction at a wedge shape hill showing rays reflected from the ground on either side of the hill, and from the hill itself.

is responsible for communication beyond the horizon at high frequencies. High in the atmosphere, solar radiation creates free electrons that reflect low-frequency waves (below about 6 MHz), allowing propagation to distances far beyond the horizon.

A. Radio Wave Absorption in the Atmosphere

Aside from condensed water, in the form of rain and clouds, the gases that are primarily responsible for radio wave absorption are oxygen and water vapor (Crane, 1996). The specific absorption in deci-Bells per kilometer due to oxygen and water vapor is plotted as a function of frequency at sea level in Fig. 7. Water vapor has absorption peaks at 22.2 GHz, as seen in Fig. 7, and at frequencies above 100 GHz. Several resonance lines are together responsible for the strong oxygen absorption peak around 60 GHz. Between the absorption peaks there are bands of lower absorption called atmospheric windows.

Over terrestrial links, the clear air absorption will be the product of the path length and specific absorption. On satellite links, the variation of absorption with altitude must be taken into account. The absorption is proportional

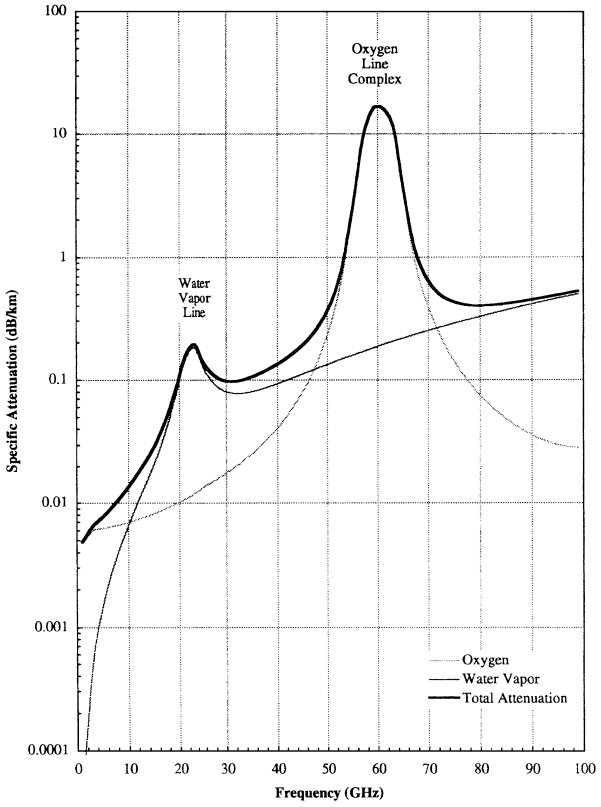


FIGURE 7 Specific attenuation in clear air as a function of frequency (From Crane, R. K. (1996). "Electromagnetic Wave Propagation Through Rain," Wiley, New York.)

to the density of the absorbing molecules, which show an approximately exponential decrease with elevation up to about 90 km. The scale height of the atmosphere is defined as the height of a column of air having uniform density equal to that at the ground, and having the same total number of molecules as would be found in a vertical column of the same area in the actual atmosphere. Since the scale height of the atmosphere is about 8 km (Hall and Barclay, 1989), clear air attenuation for overhead satellites is therefore approximately eight times the specific attenuation in Fig. 7. The attenuation would be even greater for slant paths.

The influence of water vapor and condensed water on propagation is due to the rotational motion of its polar molecules in response to an applied field. This response results in a high dielectric constant and energy absorption. Snow and ice, which have low dielectric constants because the molecules of HO are locked in a crystalline structure, do not influence radio propagation for frequencies below 30 GHz. The attenuation effect of rain and clouds on radio links is influenced by a variety of statistical parameters, including (1) the drop size distribution; (2) the density of drops, which is related to the rain rate; (3) size of the cloud or rain cell; and (4) path orientation through the cloud or rain cell.

In clouds, the droplets are very small, having radius on the order of 10 μm, so that the attenuation increases as the square of the frequency. For a cloud at 0°C having 1 g/m^3 of water, the specific attenuation at 3 GHz is 0.009 dB/km, while at 30 GHz it is 0.77 dB/km (Crane, 1996). In rain, the probability of drop size approximately follows a gamma-type distribution with the scale size of the drop radius on the order of 1 mm. The scale size increases with the rain rate, which is measured in millimeters per hour. The specific attenuation computed as a function of frequency is plotted in Fig. 8. The attenuation increases with frequency, and increases significantly with rain rate, but is small below 4 GHz.

Modeling attenuation due to rain must account for the statistical properties of the spatial distribution of rain during a rain event, as well as the statistics of the occurrence and duration of rain events. Radar measurements indicate that a rain event may be comprised of a cluster of convection cells, each of which has dimensions on the order of 1 km and duration on the order of 10 min (Crane, 1996), and the rain rate in the cell can vary with elevation. This structure is superimposed on the average rain rate within an event. The occurrence and duration of rain events, and their average rain rate, are themselves statistical properties that vary from region to region over the surface of the earth. For simplicity, all of these effects may be treated together by modeling the percentage of time that a radio link will experience attenuation above a particular value.

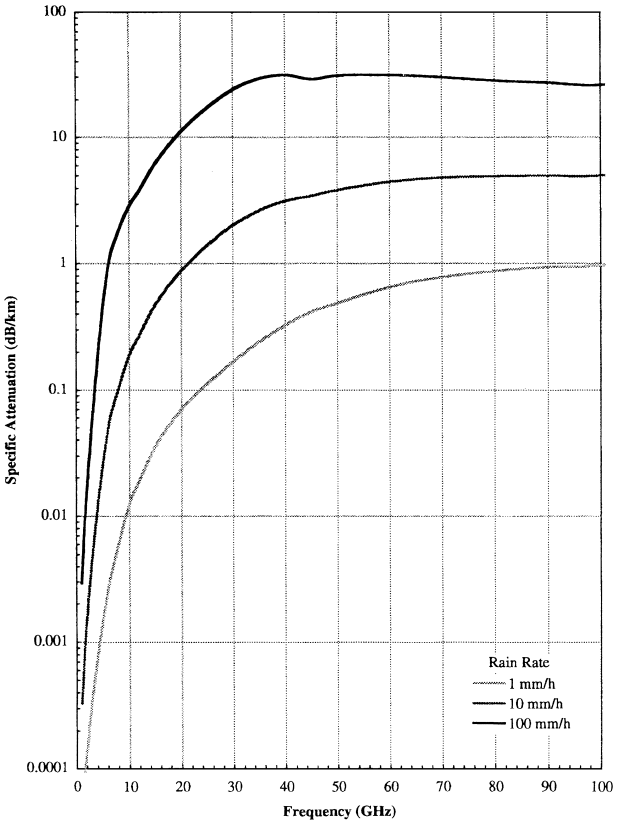


FIGURE 8 Specific attenuation through rain as a function of frequency (From Crane, R. K. (1996). "Electromagnetic Wave Propagation Through Rain," Wiley, New York.)

Such models allow the prediction of percentage of time that a radio link will be out of service.

B. Influence of the Background Density Variation

Gravity causes the density of the atmosphere to have an approximately exponential decrease with altitude. As a result, the polarizability χ of the atmosphere decreases exponentially with altitude. Because $\chi \ll 1$, the refractive index $n = \sqrt{1+\chi}$ is never much larger than unity, although it can still cause significant bending of rays propagating nearly horizontal to the earth over long paths. For convenience in describing the height variation of n, the quantity $N = (n-1)10^6$ has been introduced, and is referred to as "N-units". A plot of N versus height z above sea level is shown in Fig. 9 for the standard exponential atmospheric variation. From the value of N at sea level, $n \approx 1.0003$ there, and the $1/e$ scale height is about 7 km. A linear fit to the lower part of the curve provides a good approximation for distance up to 2 km. For the standard atmosphere, representing a worldwide average, the linear fit is $N = 289 - 39h$, where h is the altitude in kilometers above sea level.

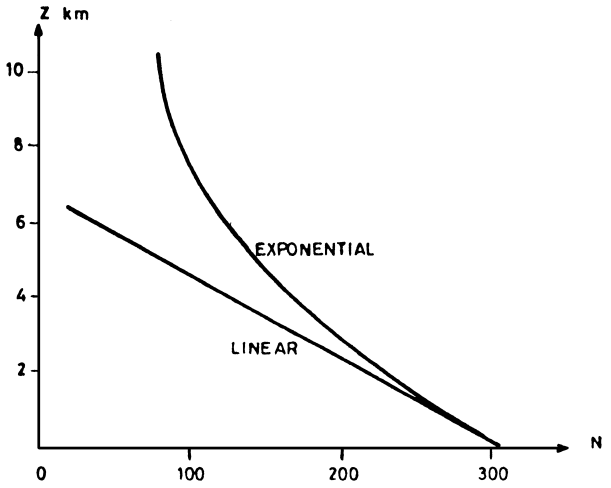


FIGURE 9 Variation of N with altitude in the standard troposphere, showing the linear fit for low altitudes (From Hall, M. P. M., and Barclay, L. W. (eds.) (1989). "Radiowave Propagation," Peter Peregrinus, Ltd.).

To satisfy Snell's law, rays going one layer to a second layer with smaller index of refraction n must bend away from the normal. Conversely, ray going into a layer with larger n must bent toward the normal. Because the basic variation of n decreases with height, rays going from one layer in the atmosphere to the next are bent downward, as shown in Fig. 10(a). For nearly horizontal propagation, the radius of curvature ρ of the ray path is

$$\frac{1}{\rho} = -\frac{1}{n}\frac{dn}{dh}. \tag{7}$$

For the standard atmosphere $\rho = 10^6/39$ or 25,600 km, which is greater than the earth's radius $a = 6370$ km. Thus bending in the standard atmosphere is not sufficient for the rays to follow the curvature of the earth. An alternative to

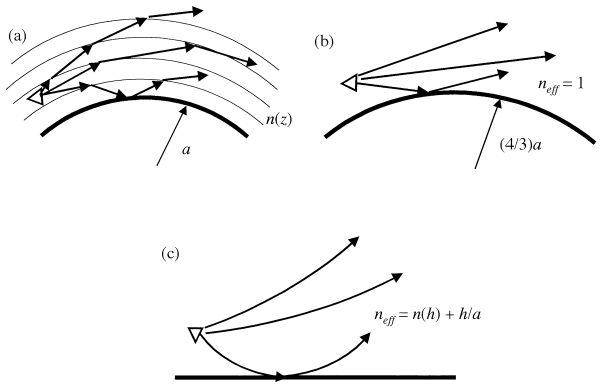


FIGURE 10 Rays in a layered atmosphere above the earth for the equivalent conditions of (a) $n(h)$ for standard atmosphere and the actual earth radius $a \approx 6370$ km; (b) uniform effective index of refraction above an earth of radius $(4/3)a$; and (c) an effective index $n(h) + h/a$ above a flat earth.

tracing the rays when n varies with altitude is to treat the atmosphere as homogeneous by using an effective earth radius a_e that is greater than a. The curvature of the effective earth $1/a_e$ is the difference between the earth curvature $1/a$ and the ray curvature $1/\rho$. With the value of ρ for the standard atmosphere, $a_e = 8{,}480$ km, which is approximately $(4/3)a$. Surface terrain features can be directly plotted on the equivalent earth surface to determine reflection and diffraction of the ray paths, which are now straight lines, as shown in Fig. 10(b).

The transformation process can be carried a step further to make the equivalent earth flat, as shown in Fig. 10(c). In this case the effective index of refraction is given by $n + h/a$, which increases with height so that the rays are curved upward, as shown. To describe the effective index of refraction, "M-units" defined by $M = (n - 1 + h/a)10^6$ are used. For the standard atmosphere, $M = N + 157h$. This transformation is particularly useful in treating ducting.

C. Tropospheric Ducting and Scattering

Nonstandard variations of temperature and humidity with altitude give rise to abnormal variations of the index of refraction, which can in turn affect the received power. The atmosphere is subrefractive when the refractive index increases with altitude, so that the rays launched by a transmitter are bent upward. Upward ray bending reduces the distance to the effective radio horizon, and limits the range of signal coverage. In contrast, the atmosphere is superrefractive when $dN/dh \leq -157$ so that the radius of curvature of the rays is smaller than the radius of the earth. As a result of the bending, rays are returned to earth where they reflect to start another hop. Additional variations in the height dependence of the refractive index can lead to ducting, while combined vertical and horizontal variations give rise to scattering.

In a marine environment there are three distinct types of ducts caused by vertical gradients of n, namely, elevated ducts, surface ducts, and evaporation ducts. The N and M profiles for an elevated duct are depicted in Fig. 11. The vertical extent of the duct is from the top of the trapping layer down to a height where M is equal to value at the top of the trapping layer. In a surface duct, the value of M at the surface of the earth is greater than at the top of the trapping layer. Propagation in these ducts is not particularly sensitive to frequency, being able to support long over-the-horizon propagation at frequencies above 100 MHz. Surface ducts occur with annual frequencies of up to 50% in areas such as the Eastern Mediterranean and Northern Indian Ocean and are responsible for most reports of extremely long over-the-horizon radar detection and communication. Effects of elevated ducts are usually limited to air-borne transmitters located close to or above the duct. Elevated ducts occur at altitudes up to about 6 km, although they are most common below 3 km.

The evaporation duct is a nearly permanent feature created by the very rapid decrease of moisture immediately above the ocean surface. The air adjacent to the ocean is saturated with water vapor and the relative humidity is thus 100%. This high relative humidity decreases rapidly with increasing height in the first few meters until an ambient value is reached that depends on the general meteorological conditions. The rapid decrease in humidity creates a trapping layer adjacent to the surface, as illustrated by the

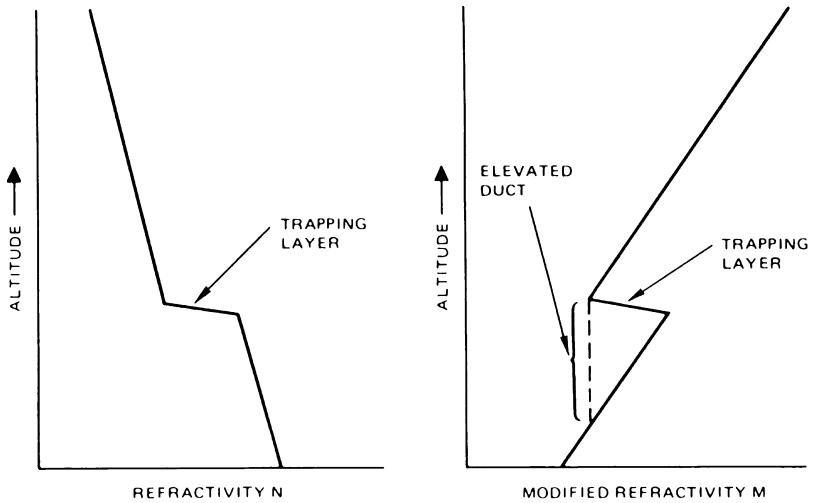


FIGURE 11 Profiles of N and M for an elevated duct. For a ground based duct M at ground level is greater than the minimum value of M at the top of the duct.

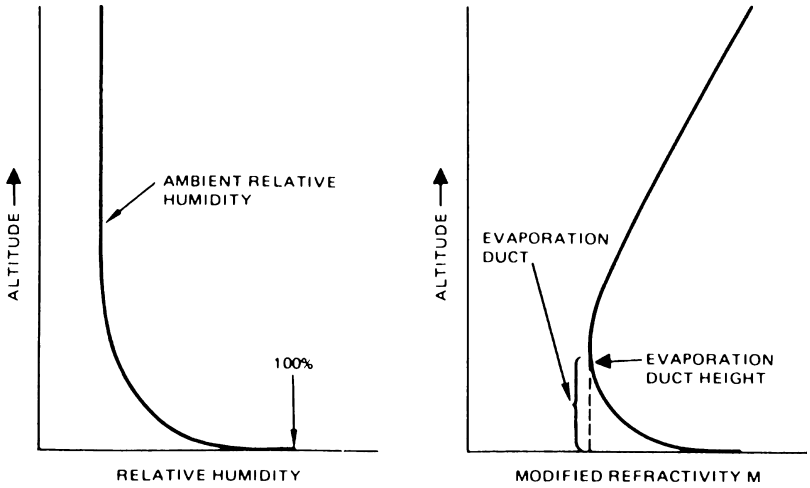


FIGURE 12 Relative humidity and *M* profile for an evaporation duct.

modified refractivity curve in Fig. 12. The height at which a minimum value of M is reached is called the evaporation duct height. Duct height is generally less than 40 m, with a long-term mean value of 8 m at northern latitudes, and up to about 30 m at tropical latitudes. Because evaporation ducts are very "leaky," they affect radio and radar terminals significantly above as well as within the duct. The frequencies that the evaporation duct can affect are strongly dependent on the duct heights, with a lower practical limit of about 3 GHz. Above this frequency the duct can give extended range for surface-to-surface communications and radar systems. The optimum frequency to achieve extended ranges appears to be around 18 GHz. At higher frequencies atmospheric attenuation and reduced reflection due to sea surface roughness begin to counteract the benefits of the duct.

Laminar flow of air mass in the troposphere can generate stratification of the atmosphere in narrow stable layers of variable size in the vertical and horizontal directions. Turbulent flow is also present in the troposphere, and generates unstable eddies that span a range of sizes down to about 1 mm. This structure results in variations of the refractive index that can scatter a small portion of the energy carried by radio waves to other directions, and can cause amplitude and phase variations in the passing wave, which are perceived as scintillation of the received power. For communication or Radar systems operating above about 300 MHz and using directive beam antennas, the scattering makes possible operation beyond the horizon. The importance of scattering for 5 GHz transmission is suggested in Fig. 13. At distances out to about 20 km, ground reflection gives interference effects. Between 20 and 40 km the waves are diffracted past earth's surface. Beyond about 40 km, tropospheric scattering is responsible for the received power.

D. Influence of the Ionosphere

In the atmosphere above about 70 km, ultraviolet radiation from the sun ionizes molecules generating free electrons, which remain unattached for considerable time because of the low density of the atoms. The electrons, and to a much lesser extent the heavier ions, can move in response to the electric field of a radio wave, thereby resulting in a current flow. Electron collisions with the neutral molecules heat the gas, which can result in a nonlinear response to high power waves. The physics of the wave–electron interaction is complicated by the presence of the earth's magnetic field, which causes the electrons to gyrate around the field lines resulting in rotation of the plane of polarization of a radio wave.

Neglecting the effect of the earth's magnetic field, the influence of the free electrons on waves having a weak electric field can be described through the effective dielectric constant (Budden, 1985)

$$\varepsilon = \varepsilon_o \left[1 - \frac{f_p^2 / f^2}{1 - j\nu/(2\pi f)} \right]. \tag{8}$$

In (8), ε_o is the dielectric constant of free space, f is the wave frequency, ν is the frequency with which electrons collide with the gas molecules. The term f_p in (8) is called the plasma frequency, and is found from the electron density using the formula $f_p^2 = N_e e^2/(4\pi^2 m \varepsilon_o)$, where N_e is the electron density, e is the charge of the electron, and m is the mass of the electron. For N_e in electrons/m^3, $f_p = 80.6 N_e$. The approximate expression on the far right of (8) is valid for $\nu < f$.

Since free electrons are generated by ionizing solar radiation, the exact dependence of N_e on altitude depend on the local time of day, season, and solar activity. Figure 14 shows the variation with height of electron

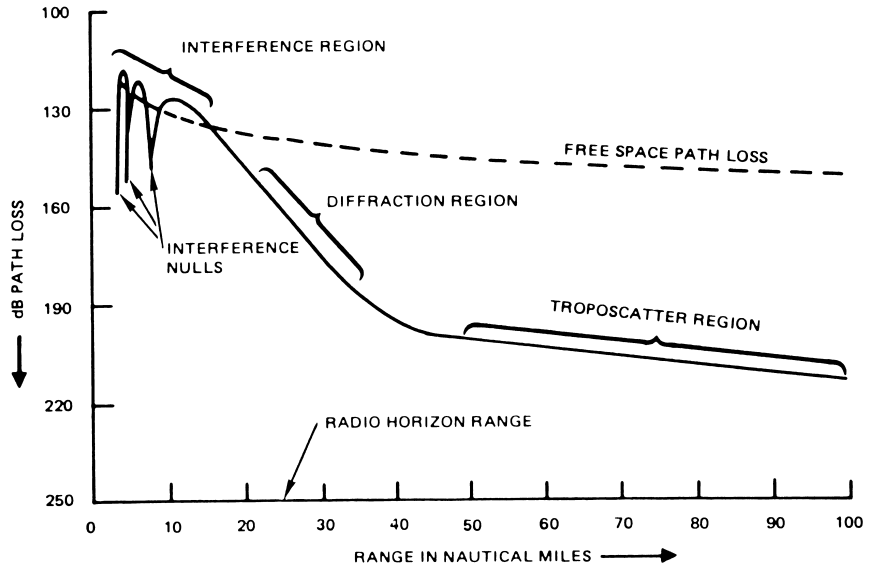


FIGURE 13 Path loss versus range illustrating the different propagation regions. The example is for terminals at 18 and 30 m elevation and $f = 5$ GHZ.

density N_e during a minimum in the sun spot cycle. Since $f_p = 8.98\sqrt{N_e}$, on a winter night the peak electron density of 10^{11} m^{-1} has a plasma frequency of 2.8 MHz. The collision frequency falls off rapidly with altitude, and $\nu < 10^5$ above 100 km. For $\nu \ll 2\pi f$ in (8), the real part ε' of the dielectric constant is given by the simple expression $\varepsilon' \approx \varepsilon_o(1 - f_p^2/f^2)$. When $f < f_p$, then ε' negative, so that the effective wavenumber $k = (2\pi f/c)\sqrt{\varepsilon'/\varepsilon_o}$ is imaginary and waves of this frequency attenuate rather than propagate. Thus low-frequency waves beamed upward into the ionosphere from the ground will propagate up to the altitude where $f = f_p$. Since they cannot propagate beyond that altitude, they will be reflected back to earth.

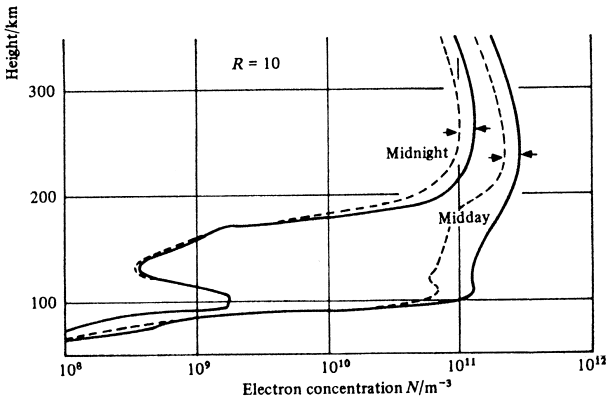


FIGURE 14 Electron density as a function of altitude in the ionosphere at a minimum of sun spots. The dashed curve is for January and the solid curve for June. (From Budden, K. G. (1985). "The Propagation of Radio Waves," Cambridge University Press.)

Waves propagating directly upward with frequencies above the maximum value of f_p are transmitted through the ionosphere with little reflection. If the frequency is greater than several times the maximum value of f_p, then ε' is close to ε_o and the radio wave propagates with less influence from the ionosphere. However, high-frequency ground-to-satellite links can experience ray bending and rotation of the plane of polarization as a result of propagation through the ionosphere.

Radio waves that are obliquely incident on the ionosphere are reflected, as shown in Fig. 15, even when f is somewhat greater than the maximum f_p. Because of the reflection from the ionosphere, waves can reach portions of the earth's surface beyond the radio horizon. The waves are subsequently reflected from the earth, after which they can again be reflected off the ionosphere. These multiple reflections allow waves to propagate around the world. At night the ionosphere is more spatially uniform,

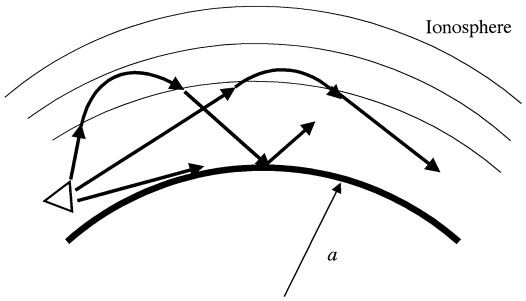


FIGURE 15 Paths of waves reflected from the ionosphere.

is effectively higher and has lower collision frequency so that the received signal is stronger. The maxim usable frequency for ionospheric reflection depends on local conditions, but is less than about 6 MHz. At very low frequencies (below 30 KHz), propagation can be viewed in terms of guided wave modes excited in the wave guide formed by the ionosphere and the earth's surface (Wait, 1996).

IV. INFLUENCE OF BUILDINGS

The influence of buildings is most strongly felt when the wavelength is smaller than the building dimensions. This condition is met for frequencies in the VHF band (30–300 MHz) and above. While the effect of buildings is significant for various broadcast services, it was the introduction of cellular mobile radio that placed the greatest emphasis on the influence of buildings. This emphasis results from two aspects of cellular networks. First, the terminals are at street level among the buildings that make up cities. More importantly, the cellular concept of frequency reuse requires not only good coverage within a cell of limited extent but also careful control of interference to nearby cells using the same frequency. The simultaneous need to provide coverage and avoid interference requires a better understanding of the propagation characteristics and greater accuracy in predicting the characteristics than are needed for broadcast services. While signal strength predictions were of primary importance for early systems, subsequent development of advanced digital radio systems created a need for additional understanding of the dispersion in time of transmitted pulses, the distribution in space of the arriving ray paths, and other higher-order channel statistics.

To support system and network design, many measurements were made in the UHF frequency range from 300 MHz to 3 GHz, and some even in the microwave and millimeter-wave spectrum up to about 40 GHz, for distances from 10s of meters out to about 20 km. For such frequencies the wavelength is less than 1 m, and hence small compared to the size of buildings. As a result, the arriving signals can be thought of as propagating along rays from the transmitter to the receiver. In urban environments there are many such paths that result from reflection and diffraction by the buildings, as well as scattering from vehicles and other obstacles in the street scene. Because of reflection, diffraction, and scattering, it is possible to receive the signal even when the antennas are not within line-of-sight (LOS) of each other.

The multiple ray paths to a street-level terminal are suggested in Fig. 16, treating the base station as transmitter. When the terminal is transmitting, reciprocity requires that the same ray paths exist with the arrows reversed. The influence of these multiple ray paths on narrow band signals and wide band pulses is discussed in the following section. In addition to the multipath effects, the average signal level will change as the terminal moves down a block, or into an intersection, where the distance to the buildings is greater. These amplitude variations are referred to as shadow fading. Finally, as the terminal moves over distances on a kilometer scale, the signal exhibits range dependence. Shadow fading and range dependence are also discussed in the following section. We complete this section with a discussion of theoretical models that explain the various observed effects.

A. Observed Multipath Effects for Narrow Band Signals

Typical procedures used for measuring narrow band or constant wave (CW) signals involve radiating a signal from a base station antenna mounted on a building, tower, or vehicle mounted mast. In most measurements the signal is received by a simple vertically polarized antenna having isotropic reception in the horizontal plane, which is mounted on a van or automobile that also contains position locating equipment and recording equipment. In this way a file or record can be generated of the mobile's

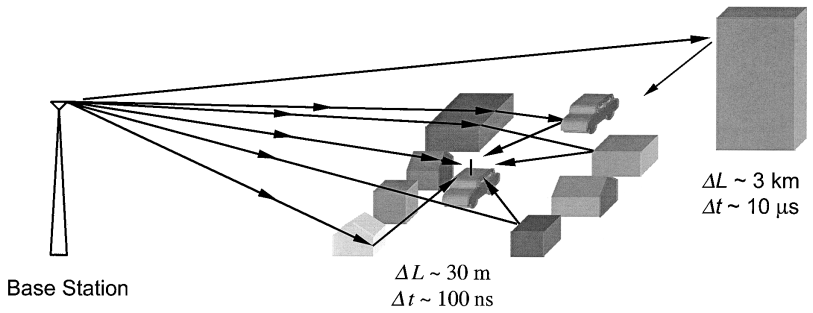


FIGURE 16 Ray paths for transmission from an elevated base station to a street level terminal showing scattering by buildings around the terminal, and scattering by distant buildings. (From Bertoni, H. L. (1999).)

position and the received signal at that position for subsequent processing.

For narrow band excitation, the received voltage envelope $V(x)$ at any point x along the street can be written as the sum of individual ray contributions, or

$$V(x) = \left| \sum V_n \exp(-jkL_n) \right|. \tag{9}$$

Here V_n is the amplitude of the arriving ray and L_n is the path length of the ray. As the terminal moves a few wavelengths λ along the street, the values of the phase kL_n in (9) will change by amounts on the order of $\pm 2\pi$, depending on the direction of the ray arrival as compared to the direction of motion. These changes in relative phase of the terms in (9) result in rapid variations of the received voltage over distances of about one half the wavelength λ. Figure 17 shows an example of this spatial variation of the received voltage amplitude measured at 910 MHz (Lecours *et al.*, 1988).

It is seen from Fig. 17 that over a distance of a few meters the signal can vary by 30 dB, which on a voltage scale is a factor of 31.6. Even over distances as small as $\lambda/2$ the signal varies by 20 dB. This small-scale variation results from the arrival of the signal at street level via multiple ray paths involving reflection, diffraction, and scattering by buildings, vehicles, and other objects in the vicinity of the terminal, as suggested in Fig. 16. The multiple rays set up an interference pattern in space through which the street level terminal moves. Since the signals arrive from all directions in the plane, fast fading will be observed for all directions of motion, e.g., a pedestrian crossing the street with a portable phone, or walking down the street.

As the terminal moves farther down the street it will pass buildings of different heights, vacant lots, street intersections, etc. In response to the variation in the nearby buildings, there will be a change in the average about which the rapid fluctuations take place, as seen in Fig. 17. This middle scale over which the signal varies, which is on the order of the building dimensions, is referred to by various authors as the shadow fading, slow fading, or log normal fading. Finally, the overall average of the signal has a systematic decrease as the distance R from the base station ranges from 100 m out to 20 km.

The fast fading can be separated from the shadow fading by averaging the envelope or magnitude of the received voltage $V(x)$ over a distance on the order of 10 m. Although the voltage is usually plotted on a deci-Bell scale, the voltage scale is used when computing the average. The result is referred to as the Small Area Average. The averaging is often done during data acquisition by taking a 1-sec average of the voltage or received power so as to reduce the amount of data that must be stored. In 1 sec a vehicle traveling 40 km/h (25 mph) covers approximately 11 m, corresponding to 33λ at 900 MHz. In this way the average in discrete small areas is obtained. Alternatively, if a complete record of the received voltage $V(x)$ is available, as in Fig. 17, the shadow fading can be separated from the fast fading using a sliding average over a window whose length is on the order of 10 m (Lecours *et al.*, 1988). The sliding average $\overline{V(x)}$, converted back into a deci-Bell scale is shown in Fig. 17. When viewed on a deci-Bell scale, the ratio $V(x)/\overline{V(x)}$ is obtained from the difference between the two curves. The difference is seen to have an irregular appearance, and is therefore described by its statistical properties.

Let $r_i = r_i = V(x_i)/\overline{V(x_i)}$ be the sampled value of the normalized voltage at a point x_i of the fast fading. Using 200 or more sampled values of the random variable r_i, a probability distribution function (PDF) can be

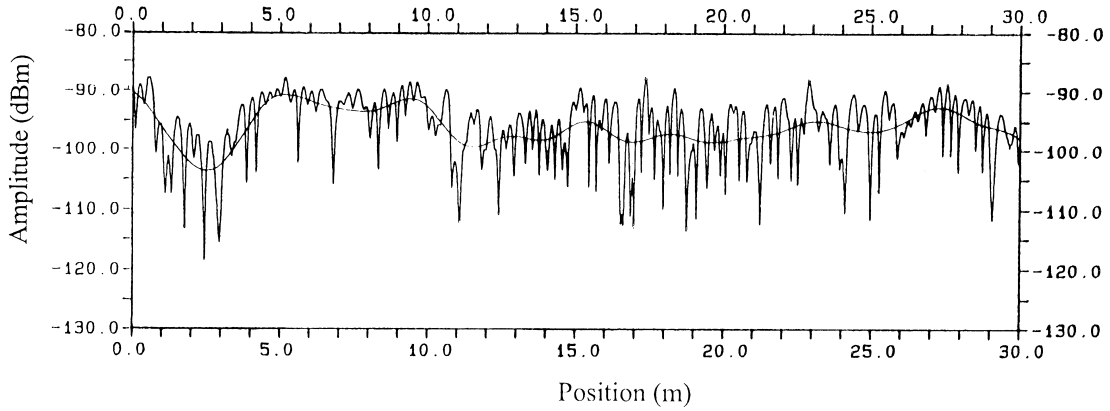


FIGURE 17 Spatial variation of a narrow band signal measured along a street showing both fast fading about the shadow fading obtained from the sliding average for non-LOS conditions (From Lecours, M., Chouinard, I. Y., Delisle, G. Y., and Roy, J. (1988). "Statistical Modeling of the Received Spinal Envelope in a Mobile Radio Channel," IEEE Trans. on Veh. Tech., Vol. VT-37, pp. 204–212.)

constructed. For locations that are heavily shadowed by surrounding buildings it is typically found that the PDF denoted by $p(r)$ approximates a Rayleigh distribution $p(r) = (\pi r/2)\exp(-\pi r^2/4)$ defined for $r > 0$. Because of this property, fast fading is often referred to as Rayleigh fading.

For locations where there is one path making a dominant contribution to the received voltage, such as when the base station is visible to the terminal, the distribution function is typically that of a Rician distribution (Parsons, 1996), whose probability density function contains an additional parameter. This parameter is often taken as the ratio K of the power carried by the strongest ray to the power carried by the weaker rays. For $K = 0$, the Rician distribution reduces to the Rayleigh distribution. For larger values of K, the fading is less pronounced than that of Fig. 17.

When the terminal is in motion, there will be a constant change in the phase of each arriving ray given by $-k(dL_n/dt)$, which appears in the receiver as a Doppler frequency shift $\Delta f = -(f/c)(dL_n/dt)$. If the terminal is traveling with velocity v, the Doppler shift for rays making an angle θ to the direction of travel will be $\Delta f = (vf/c)\cos\theta$. Considering all of the multipath ray contributions, the motion of the terminal spreads the received signal over a band of frequencies. An example of the measured Doppler spectrum of the received signal is shown in Fig. 18 (Pederson *et al*., 1997). The smooth curve labeled classical spectrum in Fig. 18 assumes a continuous and uniform distribution of arrival directions in the horizontal plane.

Because the phase of the individual ray paths in (9) depend on frequency through the term kL_i, where $k = 2\pi f/c$ the relative phases of ray paths with different lengths will vary with frequency f. As a result, the amplitude of the received voltage will also vary with frequency. An example of the frequency dependence of the radio channel about the center frequency of 910 MHz is shown in Fig. 19 for a 1.2-km link in downtown Toronto (Sousa, 1994). Accounting for the difference in the length L_i and L_j for two ray paths, the difference in phase is $(L_j - L_i)\,2\pi f/c$. For outdoor paths it is easy to conceive of scatterers giving rise to path length differences on the order of 1 km, so that the phase difference will undergo 2π phase when f changes by 0.3 MHz. In other words, the two ray contributions can go from constructive to destructive and back to constructive addition, or vice versa, in 0.3 MHz, which is consistent with the spacing between minima in Fig. 19. For indoor paths the differences in path length will be much smaller, so a greater change in frequency is required to produce the same cycle in the interference.

The path length L_n in (9) is the total distance along a ray from the base station to the various reflecting points

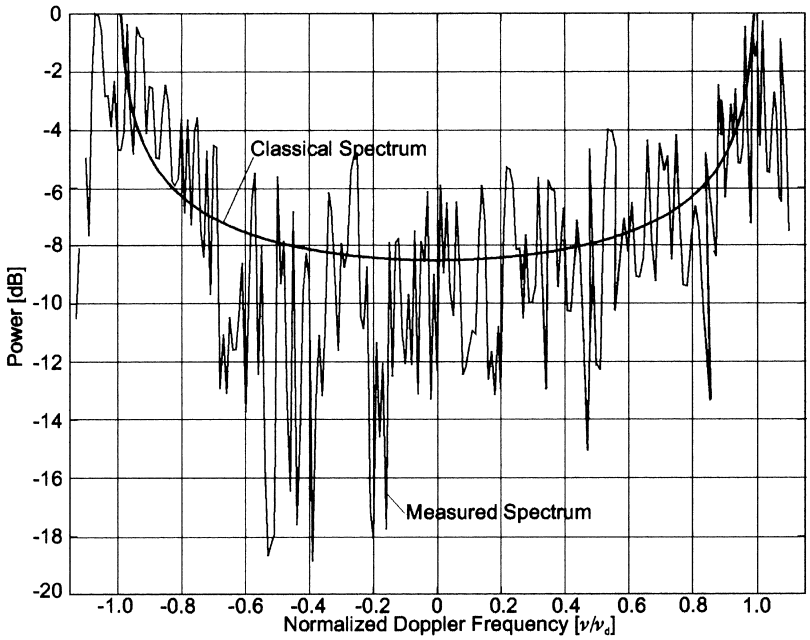


FIGURE 18 Received power spectral density due to Doppler spread at a terminal moving with velocity v. The measurements are for 1,800 MHz, and are comped with a simple theoretical prediction. The maximum Doppler shift is v/λ. (From Pedersen, K. I., Mogensen, P. E., Fleury, B. H., Frederiksen, F., Olesen, K., and Larsen, S. L. (1997). "Analysis of Time, Azimuth and Doppler Dispersion in Outdoor Radio Channels," Proc ACTS.)

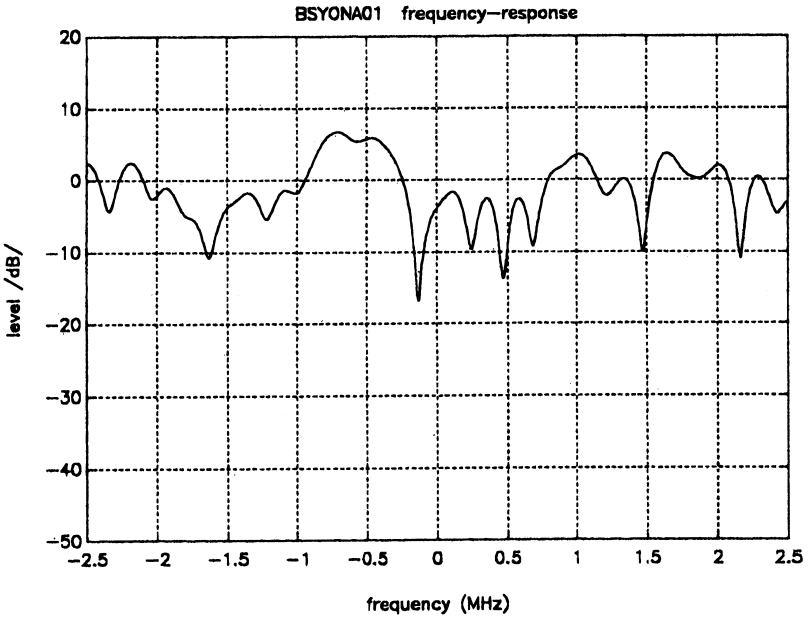


FIGURE 19 Frequency variation about the center frequency of 910 MHz for a 1.2-km radio link in downtown Toronto. (From Sousa, E. S., Jovanovic, V. M., and Daigneault, C. (1994). "Delay Spread Measurements for the Digital Cellular Channel in Toronto," IEEE Trans. on Veh. Tech, Vol. VT-43, pp. 837–847.)

and/or scatterers, and then to the terminal. The motion of scattering objects, such as vehicles, pedestrians, and trees, causes a change in L_n even when the terminal is stationary. The resulting change in the phases of the individual rays will therefore change with time, causing the amplitude of the received signal to change over time intervals on the order of 100 msec.

Some of the scattering processes that are responsible for the multipath arrivals will depolarize the arriving fields. Measurements have been made of the co-polarized (vertical) and cross-polarized (horizontal) electric fields received at base stations for vertically polarized radiation from a mobile. Both polarizations exhibit fast fading as the mobile moves along the street, with fading patterns that are uncorrelated. The ratio of the average power carried by the cross-polarized field to that of the co-polarized field is around -7 dB in cities and -13 dB in suburban areas. In the case of hand-held terminals, where the inclination of the antenna is significant, one can expect more nearly equal average power to be received at the base station in both the vertical and the horizontal polarizations.

B. Observed Multipath Effects for Pulsed Signals

The transmission of digital signals is influenced by the multipath character of the channel through the echoes that spread the received pulse over time. Existing and proposed telecommunication systems employ a limited bandwidth for each channel, which is typically 1–5 MHz wide, although some concepts have discussed far wider bandwidths. For center frequency in the UHF band, such radiated pulses have a small fractional bandwidth, and hence each pulse contains many RF cycles. As a result, the finite bandwidth channel response is important, rather than the infinite bandwidth impulse response.

We can generalize expression (9) to account for the pulse nature of the transmission. Let $w(t)\exp(jwt)$ be the voltage, normalized to the free space path loss, that would be received when the transmitter and receiver are in free space with a small separation to avoid time delays. When placed in a multipath environment, the received voltage will be $V(x,t)e^{j\omega t}$, where $V(x,t)$ is the complex amplitude. Neglecting distortion of the individual pulses as a result of reflection, diffraction or scattering, the complex received voltage amplitude is found by summing the contributions from the individual ray paths as

$$V(x,t) = \sum w[t-\tau_n(x)]V_n(x)\exp(-jkL_n). \qquad (10)$$

Here $\tau_n(x)$ is ray time delay given by L_{ni}/c, where L_n is the ray path length and $V_n(x)$ is the ray amplitude.

At any location x, the received voltage envelope is $|V(x,t)|$, while the envelope power $P(x,t)$ is proportional to $|V(x,t)|^2$. The envelope power plotted as a function of time is called the power delay profile. One example of a

power delay profile measured at 850 MHz using a channel sounder with effective pulse width of 50 nsec is shown in Fig. 20 (Devasirvatham, 1988). Significant power is received out 2 μsec, which is much larger than the duration of the original pulse. The signal consists of three clusters, or groups of arriving rays, each of which is wider than the original pulse. Within each cluster the signal decreases in amplitude, and subsequent clusters have smaller amplitudes.

The clusters of Fig. 20 can be explained by referring to Fig. 16. Buildings and other objects near the terminal represent one class of scatterers producing the multipath arrivals in Fig. 16. Since the dimensions of a city block are on the order of 100 m, these nearby objects can produce signals delayed by around 0.3 μsec. Large buildings and structures at greater distances can cause stronger scattered signals that arrive with delays greater than 1 μsec. Signals scattered from large buildings will be rescattered by objects nearer to the terminal, resulting in a cascade of scattered arrivals at the terminal. Each cluster of arriving rays exhibits decreasing signal amplitude and lasts a few tenths of a microsecond as shown in Fig. 20, while the individual clusters arrive with larger time delay, and the cluster amplitudes decrease with increasing time delay. Figure 20 has been selected for the clarity with which it shows the ray clusters. In most power delay profiles the amplitudes of the later clusters are smaller, and are therefore harder to distinguish.

The parameter RMS delay spread τ_{RMS} is commonly used to quantify the spread in time of the received signal. Delay spread is from the received voltage $V(x,t)$ using the expressions

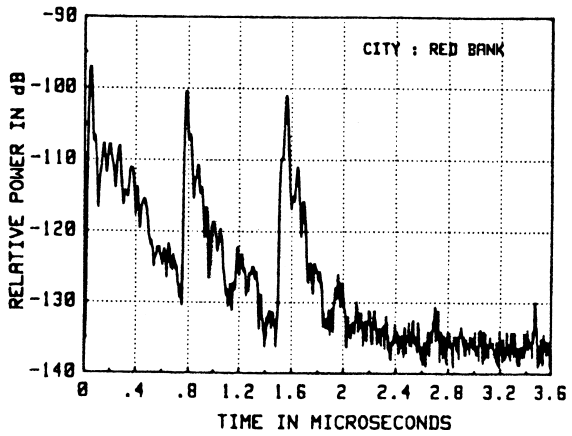


FIGURE 20 Power delay profile measured in Red Bank, NJ, using a 850-MHz channel sounder with the equivalent of a 50-nsec pulse width. (From Devasirvatham, D. M. J. (1988). "Radio Propagation Studies in a Small City for Universal Portable Communications," Proc. of the IEEE Veh. Tech. Conference, pp. 100–104.)

$$\tau_{RMS}^2 = \frac{\int_0^\infty (t - T_o)^2 |V(x,t)|^2 \, dt}{\int_0^\infty |V(x,t)|^2 \, dt},$$

where

$$T_0 = \frac{\int_0^\infty t |V(x,t)|^2 \, dt}{\int_0^\infty |V(x,t)|^2 \, dt}. \tag{11}$$

Measurements in various cities show that τ_{RMS} varies randomly from one location to the next. The statistical properties of τ_{RMS} are similar in the different cities, and appear to be only weakly dependent on base station antenna height or frequency, although the mean value exhibits a systematic increase with distance R from the base station. These results are summarized in the Greenstein model $\tau_{RMS} = \tau_0 \sqrt{R} \Psi$, where τ_0 is the mean value of τ_{RMS} at a distance of 1 km from the base station, and has value ranging from 0.4 to 1.0 μsec (Greenstein *et al.*, 1997). Also, R is in kilometers and Ψ is a random variable such that $10 \log \Psi$ has a Gaussian distribution with standard deviation between 2 and 6.

Other higher-order propagation parameters are important in designing radio systems that can improve link quality and increase network capacity. For example, designing new radio systems that use multiple antennas requires information on the angle spread of the ray paths. As suggested in Fig. 16, and found from measurements, the rays at the mobile take all directions about the mobile in the horizontal plane, and up to 30–40° in the vertical plane. At the base station, the rays lie in a limited range of angles in the horizontal plane. The RMS angle spread θ_{RMS} is used to quantify the angle spread at the base station. The angle spread is defined by expressions similar to (11) with time replaced by angle of arrival. Like delay spread, angle spread varies randomly from one location to another, and is therefore treated as a random variable. The cumulative distribution functions of the random variations for different base station antenna heights are shown in Fig. 19. The distribution is strongly dependent on antenna height, with the median value increasing from 5 to 10° as the antenna is lowered from somewhat above to somewhat below the building heights (Cheon *et al.*, 2000).

C. Shadow Fading and Range Dependence

When the fast fading is averaged over a small area, as in Fig. 17, there remains a slower random variation from one small area to the next that results from differences in shadowing by the buildings that are present along the radio path. When making measurements of received signal, the

drive path typically meanders from within 100 m from the base station out to distances of 20 km or more. The results of one such measurement is shown in Fig. 21, where the small area average power P_{dB} is plotted in a deci-Bell, the distance R from the base station on a logarithmic scale (Ott and Plitkins, 1978). In the plot each dot represents the small area average power $U(R_i)$ in a small area located at a distance R_I from the base station.

The dots in Fig. 21 cluster along a straight line, whose equation is of the form

$$P_{dB}(R) = 10\log P_T + 10\log A - n10\log R, \qquad (12)$$

where P_T is the transmitted power. When converted into an expression for the power in watts, (12) becomes $P = P_T A/R^n$, where n is the slope index and A is the intercept at $R = 1$ in the distance units of the measurements. The straight line in Fig. 20 represents the average range dependence of the signal, and the shadow fading corresponds to the deviation $U(R_i)$–$P_{dB}(R_i)$ of the small area average from the average range dependence.

The distribution of the shadow fading, as defined here, is nearly a Gaussian independent of the location of the measurements. Examples of the cumulative distribution functions (CDF) for 955 and 1845 MHz obtained from simultaneous measurements made at the two frequencies on a drive route in Copenhagen (Mogensen *et al.*, 1991) are shown in Fig. 22, where a Gaussian scale is used for

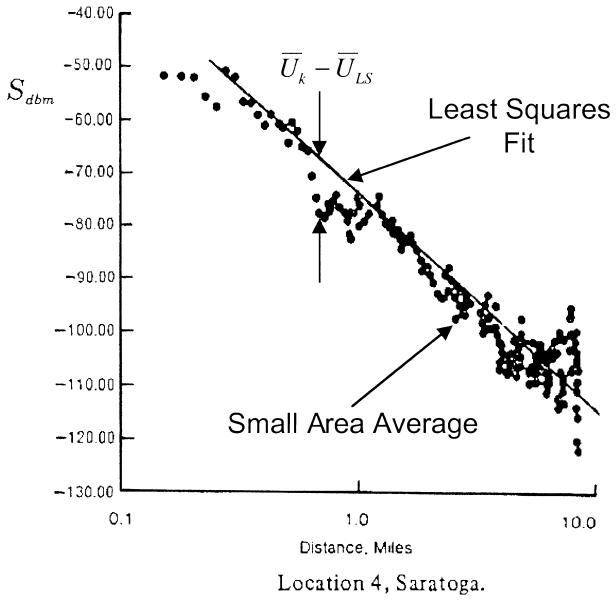


FIGURE 21 The least-square fit line to the small area averages measured on flat terrain gives the range dependence, while the deviations from the least squares fit line are the slow fading. Small area averages are based on Ott, G. D., and Plitkins, A. (1978). "Urban Path-loss Characteristics at 820 MHz," IEEE Trans. Veh. Tech., Vol. VT-27, pp. 189–197.

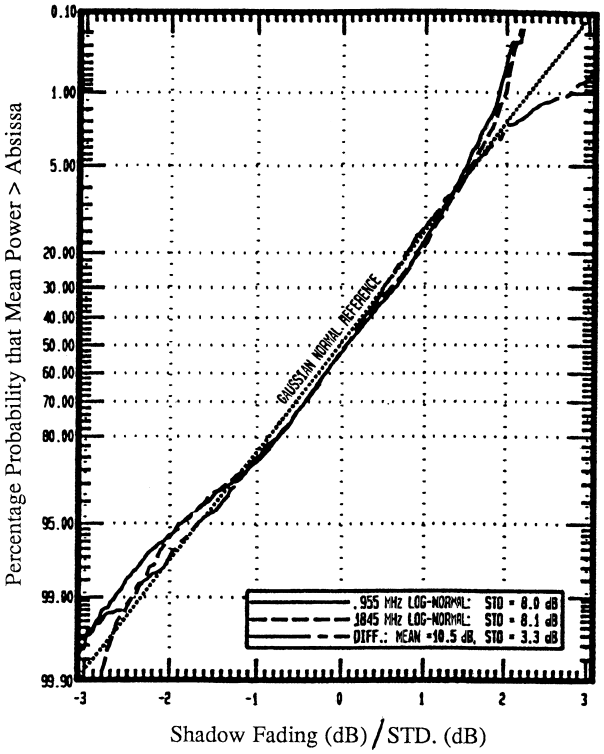


FIGURE 22 Cumulative distribution functions plotted on a Gaussian scale of the slow fading are measured simultaneously at 955 and 1845 MHz in Copenhagen. Also shown are the ideal Gaussian distribution and the distribution of the difference in slow fading at the two frequencies. (From Mogensen, P. E., Eggers, P., Jensen, C., and Andersen, J. B. (1991). "Urban Area Radio Propagation Measurements at 955 and 1845 MHz for Small and Micro Cells," Proc. of GLOBECOM'91, Phoenix, AZ, pp. 1297–1302. © IEEE.).

the vertical axis. With this distorted scale, the CDF for a Gaussian distribution plots as a straight line, as indicated by the dotted line. In this plot the random variable U for each frequency is normalized to its standard deviation σ_{SF}, which is 8.0 at 955 MHz and 8.1 at 1845 MHz. The vertical scale gives the percentage of points at which the received power is greater than or equal to the value indicated on the horizontal axis. For example, at 1845 MHz about 20% of the measurement points receive power greater than the mean power in deci-Bells plus 1 SD $\sigma_{SF} = 8.1$ dB. We conclude from Fig. 22 that the distribution function for the shadow fading follows a Gaussian distribution, except at high values of received power.

The observation that $U(R_i)$–$P_{dB}(R_i)$ is approximately Gaussian in any city can be explained by recognizing that the ray paths undergo sequential diffraction processes at buildings along the path. The irregular nature of the buildings imparts random amplitude variations to the signal. When expressed on a deci-Bell scale, the random processes are additive, and by the central limit theorem, the

sum of random processes tends to a Gaussian distribution, independent of the exact distribution function of each random process.

At each point used for the measurements leading to Fig. 22, the difference between the received power at 955 and 1845 MHz was obtained for the same transmitted power. On average the power at 1845 MHz was 10.5 dB lower than the received power at 955 MHz. The cumulative distribution of this difference is also shown in Fig. 22, where it is normalized to its standard deviation of 3.3 dB. This low value of standard deviation of the difference in received power indicates that the shadow fading at the two frequencies is nearly the same. As in the case of fast fading, other statistical parameters of the shadow fading are important to the system designer. One such parameter is the degree of correlation between the shadow fading for the links from the terminal to two different base stations, which influences the interference experiences from other cells. Another parameter is the correlation length of the shadow fading, which ranges from 5 m in urban environments to 300 m in the suburbs.

Measurements leading to slope-intercept models of the type indicated in (12) have been made in many cities. One set of measurements that has been widely used in the development of cellular systems was made in Tokyo by Okumura *et al.* (1968) for high base station antennas ($30 \leq h_{BS} \leq 200$ m) over a wide frequency range ($150 \leq f \leq 1500$ MHz). Hata (1980) subsequently converted the graphical presentation into simple formulas for distances in the range $1 \leq R \leq 20$ km that make it easy to compute the path loss. For example, the slope index n in (12) in the Hata model is given by $4.49–0.655 \log h_{BS}$, which ranges in value from $n = 3.52$ at $h_{BS} = 30$ m to $n = 2.98$ at $h_{BS} = 200$ m. If the base station antenna height is reduced to $h_{BS} = 20$ m, which is nearer to the height used in urban areas but below the range of the measurements, then $n = 3.64$. Other measurements for low base station antennas and distances in the range 100 m $\leq R \leq 2$ km have found similar slope index models to hold for the propagation. The slope index n increases further as the base station antenna is brought down to rooftop level, and can even be greater than 4 when the base station antenna is below the rooftops. In the next section we will give a theoretical explanation for the dependence of n and A on the frequency and link geometry.

D. Ray Tracing Models for Site-Specific Predictions

A city is identified by its tall buildings and other unique features that usually occupy only a small portion of any metropolitan area. Away from the high-rise core, cities are composed of rows of buildings having nearly uniform height, with occasional high buildings. The buildings are taller and closer together near the core, becoming lower and more spread out near the edge of the city and into its suburbs, and finally turning into isolated houses and other buildings in the far suburbs and surrounding rural regions. Except for the distant suburbs and rural regions, the spacing between neighboring houses is less than the width of the houses. In some cities the neighboring buildings are attached to each other with no gap between them. The houses are thus organized in rows along the streets, whose presence imposes a significant spacing between the fronts of the houses. Likewise, there is ample back-to-back spacing to accommodate a yard.

In the high-rise core, and other areas where the building features are not regular, propagation characteristics can be made by ray tracing using a database of the specific building shapes. As opposed to ray tracing for other applications, ray tracing among buildings is complicated by the many reflections and diffractions along the ray paths that must be accounted for. On the other hand, the fact that the faces of buildings are almost always vertical and flat simplifies the ray code.

Ray tracing codes based on the three-dimensional pin cushion method have been written, but are limited by running time to a total of two diffraction events along a ray path. However, if all of the building walls are vertical, then a faster two-dimensional pin cushion method is adequate to find the ray paths in the horizontal plane. The vertical component of the paths for rays reflected from the walls and diffracted at vertical corners can then be found analytically since all segments of the ray between the antennas, or between antennas and the ground reflection point, have the same slope. While this approach can include rays that travel over buildings, it requires a further approximation to include rays diffracted at horizontal edges. In order to include these rays, it is necessary to assume that the ray segments leaving the horizontal edge lie in the vertical plane of the incident ray, or in the vertical plane of the reflected ray. Placing the diffracted rays in the vertical planes, rather than on the cone whose half angle is equal to the angle between the incident ray and the edge, distorts the subsequent ray path. The distortion is small when the horizontal edge is close to being perpendicular to the vertical plane containing the incident ray, or the slope of the incident and diffracted rays in the vertical plane is small.

We call the foregoing approach the vertical plane launch (VPL) method (Liang and Bertoni, 1998). The base station is treated as launching a series of these vertical planes that are allowed to go over buildings, as well as being reflected at vertical walls, and diffracted at vertical corners. Several vertical planes are shown in Fig. 23 for the buildings of the high-rise section of Rosslyn, VA. Unfolding the vertical planes, the vertical trajectory of the rays can be found

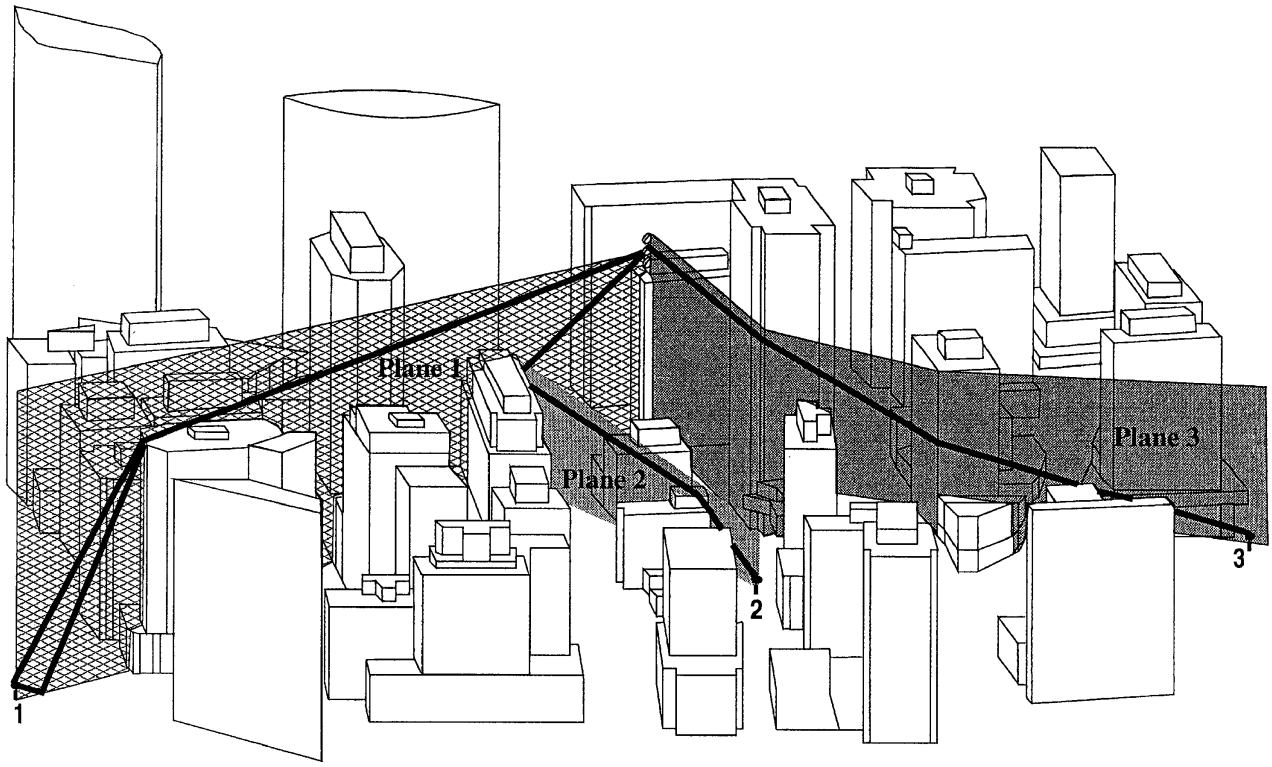


FIGURE 23 A perspective view of the buildings in Rosslyn, VA showing the vertical planes launched from a rooftop base station, and the ray paths in those planes. (From Liang, G., and Bertoni, H. L. (1998). "A New Approach to 3-D Ray Tracing for Propagation Prediction in Cities," IEEE Trans. on Antennas and Propagation," Vol. AP-46, pp. 853–863.)

analytically, thereby giving the ray paths in three dimension shown in Fig. 23. When viewed from above, the vertical planes launched by the source appear as rays in a two-dimensional pin cushion approach. However, because the planes can go over the buildings, the rays now form a binary tree with branching at the exterior surfaces of the buildings. Vertical corners that are illuminated act as secondary sources, whose rays are traced by again using the pin cushion method. Compared to three-dimensional ray tracing, each vertical corner initiates only a single two-dimensional ray trace, while diffraction at horizontal corners is treated analytically without requiring additional ray tracing. As a result, the VPL method can account for double diffraction at vertical edges, in addition to multiple diffraction over buildings.

Predictions made using the VPL method are compared in Fig. 24 with measurements at 1900 MHz made in Rosslyn. In Fig. 23 the antenna is situated atop a building at a height $h_{BS} = 42$ m, and points south (out of the page) with a 60° beam width and 6° down tilt. The predictions and measurements are plotted versus distance along the street starting from right rear in Fig. 23, just behind the first row of buildings and ending near Point 1. Typically,

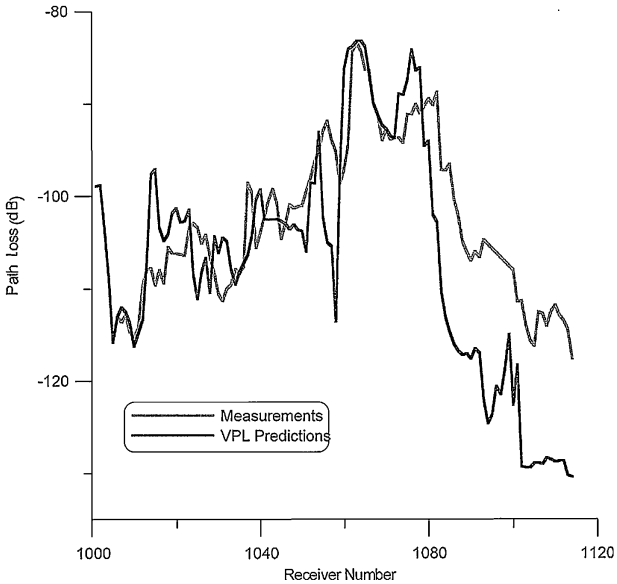


FIGURE 24 Comparison of measured and predicted path gain observed along a street through Rosslyn, VA for the rooftop base station of Fig. 23 operating at 900 MHz. The distance between receiver points is 5 m.

ray tracing predictions follow the measurements, but give greater variations of the received signal. For all base station locations, the prediction errors typically had an average of a few deci-Bells and a SD of 7–10 dB (Liang and Bertoni, 1998).

As another example of the use of the ray methods for simulating channel statistics, we have computed the RMS delay spread and RMS angle spread as seen at the base station for mobiles located in a section of Munich. For the simulations the base station antenna is located at a height of $h_{BS} = 74$ m, which is well above the tallest building, and several hundred receivers are distributed randomly over the streets. The delay and angle spreads are plotted versus R as the individual points in Fig. 25. Regression fit lines to the spreads are shown as solid lines in Fig. 25, which show that delay spread increases with distance from the base station, while angle spread does not. The dashed curve in the delay spread plot represents the Greenstein model discussed in Section IV.B and agrees overall with the simulations. There is considerable random variation of the delay spread and angle spread about the regression fit lines. These variations have a lognormal variation, as found from measurements.

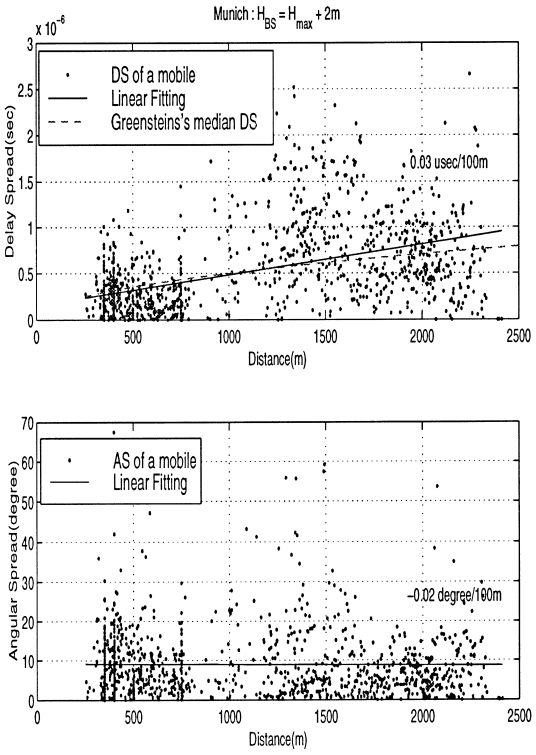


FIGURE 25 Scatter plots of simulated RMS delay and angle spread versus distance in Munich. The solid lines are obtained from least squares fits to the delay and angle spread. (From Cheon, C., Bertoni, H. L., and Liang, G. (2000). "Monte Carlo Simulation of Delay and Angle Spread in Different Building Environments," IEEE Veh. Tech. Conference.)

E. Path Gain Predictions in Low-Rise Environments

As noted previously, away from the high-rise core of a city, the buildings are nearly equal in height and organized in rows along the streets with a significant spacing between the fronts of the buildings and a nearly equal spacing between the backs. When computing the range dependence, the propagation must take place over the rows of buildings, since significant attenuation is associated with propagation through the buildings. Assuming the buildings are all the same height and ignoring the gaps between them leads to the representation of the rows of buildings shown in Fig. 26. The separation between rows ranges from about 40 m in older sections of cities, to 60 m or more in recent developments. As a result, rays from a base station to a mobile at a distance of 1 km must travel over approximately 20 rows, depending on the direction relative to the street grid. Because the rays propagate at a small angle to the horizontal, the ray fields reaching the rooftops near the mobile are reduced by the presence of the previous rows. Finally, the fields at the rooftops closest to the mobile are diffracted down to street level along several paths, two of which are indicated in Fig. 26.

The path gain for propagation from an elevated base station antenna to a terminal is therefore the product of three factors: (a) the free space path gain PG_0, (b) the reduction Q^2 in the fields arriving at the buildings near the mobile due to diffraction past the previous rows of buildings, and, (c) the reduction Λ^2 in the fields as they diffract down to ground level. When expressed in dB, the path gain is

$$PG_{dB} = 10 \log PG_0 + 10 \log Q^2 + 10 \log \Lambda^2. \quad (13)$$

The free space path gain is given in (2) for the horizontal separation R. The term Λ^2 is found in closed form using diffraction theory for a single edge. However, Q^2 is more difficult to evaluate since the diffraction takes place past many rows, and each diffracting edge lies in the transition region of the previous edge (Bertoni, 2000).

In Fig. 26 the ray diffracted at the rooftop just before the mobile has the strongest field. Other diffracted rays contribute nearly equally to the small area average power, so that doubling the contribution from the strongest ray yields the average: for R much larger than the separation between the rows, and using the diffraction coefficient $D(\theta)$ for an absorbing edge (Felsen and Marcuvitz, 1973).

$$\Lambda^2 \approx \frac{2}{\rho}|D(\theta)|^2 \approx \frac{\lambda \rho}{2\pi^2 (H_B - h_m)^2}. \quad (14)$$

Here ρ is the distance from the rooftop to the mobile, θ is the angle through which the ray is diffracted, and H_B

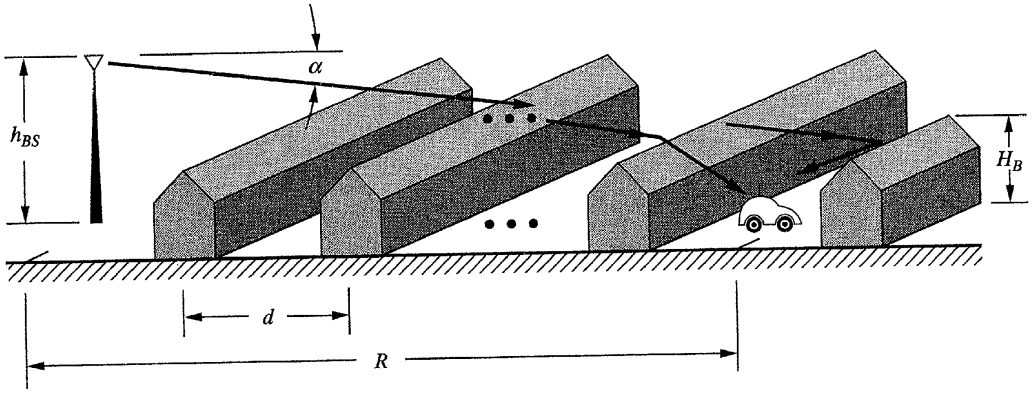


FIGURE 26 Away from the high-rise urban core, propagation takes place over the intervening rows of buildings, with the signal diffracted down to street level at the buildings next to the subscriber. (From Bertoni, H. L. (2000). "Radio Propagation for Modern Wireless Systems," Prentice-Hall PTR, Englewood Cliffs, NJ.)

is the height of the building. This simple expression gives the height gain of the mobile antenna, which is seen to depend on the relative height between the mobile and the surrounding buildings rather than the height above ground.

It is necessary to account for multiple diffraction past many edges when evaluating Q. One approach to the evaluation of Q is through the repeated application of the physical optics approximation (Bertoni, 2000). The physical optics integrals can be carried out numerically, or converted to a series of Borsma functions. For high base station antennas, the results depend on the wave length, row separation d and link geometry through the dimensionless parameter g_p given by

$$g_p = \frac{h_{BS} - H_B}{R}\sqrt{\frac{d}{\lambda}}. \qquad (15)$$

The results of the numerical evaluation of Q lie on the solid curve in Fig. 27. A polynomial fit to the numerical evaluations is

$$Q(g_p) = 3.502g_p - 3.327g_p^2 + 0.962g_p^3. \qquad (16)$$

For low base station antennas near the rooftops, g_p is small and Q is proportional to g_p, and is seen proportional to $1/R$, as seen in (15). In this case, the combined dependence of Q^2 and PG_0 on R in (13) gives the range index $n = 4$.

The dashed curve in Fig. 27 is a simple linear fit to $Q(g_p)$ over the range of g_p appropriate to cellular systems. This approximation is $Q(g_p) \approx 0.1(g_p/0.03)^{0.9}$, so that Q^2 is proportional to $1/R^{1.8}$. As a result, the range index in (13) is $n = 3.8$, which is close to the value $n = 3.64$ of the Hata model for $h_{BS} = 20$ m. In addition to the slope index, the amplitude predicted by the foregoing expression for $Q(g_p)$ and (13) agrees with measurements. The multiple diffraction model outlined above can also be used for antennas below the rooftops, and to evaluate the shadow fading when the buildings have gaps between them along the street, and are not of uniform height (Bertoni, 2000).

V. INFLUENCE OF VEGETATION

The leaves, branches, and trunks of trees act as a random medium that scatters and absorbs radio waves, resulting in significant attenuation of high-frequency radio waves. Describing the effect of vegetation over the frequency range from HF to mm waves (3 MHz to 30 GHz) is complicated because the leaves, branches, and trunks can be small compared to wavelength at low frequencies, but large compared to wavelength at highest frequencies. The different ways that trees are grouped together further complicates the description. At one extreme is a natural dense forest having continuous canopy whose top is nearly flat. In less dense forests, the tree canopy is broken by open

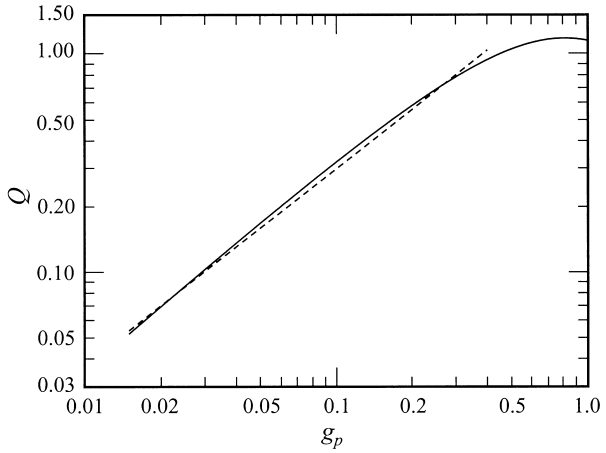


FIGURE 27 Variation of the field reduction Q with the dimensionless parameter g_p. The dashed line corresponds to the simplified approximation. (From Bertoni, H. L. (1999).)

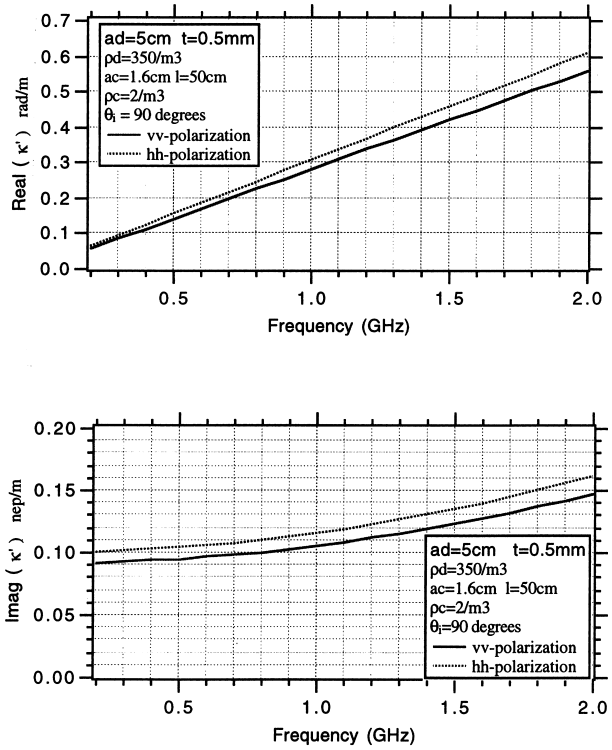


FIGURE 28 Plots of κ' and κ'' versus frequency for the foliage parameters given in Torrico, S. A. (1998). "Theoretical Modeling of Foliage Effects on Path Loss for Residential Environments," Dissertation, School of EE, The George Washington University.

areas. In cities trees are sometimes planted in rows along the streets, in small groups, or individually. For simplicity, we limit the discussion in this section to the case of a dense forest.

For ground-to-ground propagation over short distances in a forest, the tree trunks have the most significant effect on the propagation. However, for large separation the attenuation past the trunks becomes large, and other paths that are predominantly in the air give the dominant contribution. In this case, and when one of the antennas is elevated, a portion of the propagation path lies through the canopy, so that the leaves and branches are most important.

A. Propagation and Attenuation of the Coherent Field

Treating the leaves and branches or trunks as a random medium permits the theoretical evaluation of the attenuation. Inside the random medium, the total fields consist of a coherent part and a diffuse part. The coherent fields are associated with the forward scatter past the individual leaves and branches or trunks, while the diffuse fields result from scattering away from the forward direction. As the radio waves propagate into the random medium, the coherent fields are initially dominant. However, they decay relatively rapidly through absorption and scattering into the diffuse fields. After several nepers of decay of the coherent fields, the diffuse fields become dominant and continue to decay, but at a much lower rate.

For plane wave propagation, the influence of the random media on the coherent field may be described by change $\kappa = \kappa' - j\kappa''$ in the propagation constant from the free space propagation constant $k = 2\pi/\lambda$. Since the leaves and branches, or trunks, are distributed only over a limited range of angles about the vertical, the value of κ is different for vertically and horizontally polarized waves, and will depend on the propagation direction with respect to the vertical. The real and imaginary parts of κ are plotted versus frequency in Fig. 28 for both polarizations and propagation parallel to the ground through the tree canopy for a typical density of leaves and branches (Torrico, 1998). The specific attenuation α_{dB} dB/m is $8.69\kappa''$. At $f = 1$ GHz the specific attenuation is close to 1 dB/m. The change in propagation constant κ can be related to an equivalent polarizability $\chi = \chi' - j\chi''$ of the medium using the relation $\varepsilon_r \equiv 1 + \chi = (1 + \kappa/k)^2$ or $\chi \approx 2\kappa/k$. Taking the values from Fig. 28, $\chi' \approx 0.027$ for all frequencies while χ'' ranges from 0.04 to 0.01 between 200 MHz and 2 GHz.

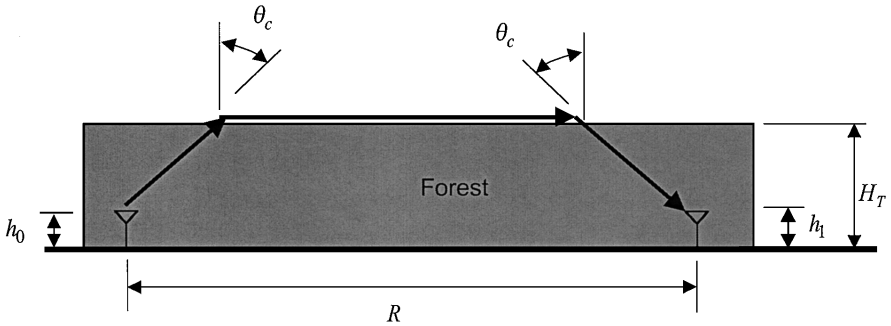


FIGURE 29 Mobile-to-mobile propagation via the lateral wave propagating above the forest canopy.

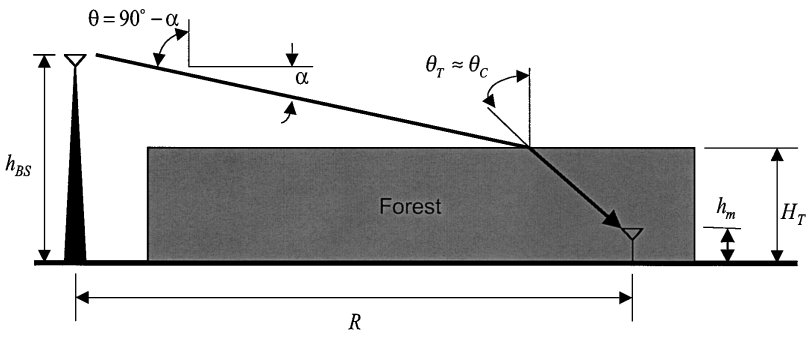


FIGURE 30 Refraction path for propagation from an elevated base station to a terminal inside a forest.

B. Path Loss to Terminals in Forests

For long mobile-to-mobile links in forests at frequencies in the 2- to 200-MHz band, Tamir (1967) showed that the propagation takes place via a lateral wave in the air propagating parallel to the forest canopy, as indicated in Fig. 29. This path starts with a ray propagating in the forest layer at the critical angle $\theta_c = \arcsin(1/\sqrt{1+\kappa'})$. The ray from the source incident on the top of the forest at the critical angle excites fields in the air propagating parallel to the tree tops. This field in turn radiates back into the forest at the critical angle. Neglecting ground reflection, and for horizontal separations R that are large compared to the vertical dimensions, the path gain PG is found to be (Tamir, 1967)

$$PG = \left(\frac{3}{2|\chi|}\right)^2 \left(\frac{\lambda}{2\pi R}\right)^4 \times \exp[2(2H_T - h_0 - h_1)k\,\mathrm{Im}\{\sqrt{\chi}\}]. \quad (17)$$

Here H_T is the height of the trees and h_0 and h_1 are the heights of the two antennas. Note that $k\,\mathrm{Im}\{\sqrt{\chi}\} < 0$ and is equal to $\mathrm{Im}\{k\sqrt{\varepsilon_r - 1}\}$, the imaginary part of the vertical component of the wave number in the forest canopy at the critical angle. In (17), the received signal decreases as $1/R^4$, corresponding to a slope index $n = 4$.

Using the Tamir model at frequencies above 100 MHz requires further research into the nature of the diffuse fields in the forest when the wave in the air is at near glancing incidence. Because the relative dielectric constant of the forest is only slightly greater than unity, the critical angle θ_c is close to 90° and the length of the critical angle ray through the trees is much longer than the height of the trees h_T. Since forest trees may be well over 10 m high, the length of the critical angle ray is longer than the distance at which the diffuse field dominates.

When one antenna is above the trees, as in Fig. 30, the path gain can be found using ray optical theory. It is necessary to account for the free space path gain to the top of the forest canopy, transmission into the forest canopy, and the attenuation from the top of the canopy down to the terminal. Taken together, the path gain for $R \gg h_{BS} - H_T$ is given by

$$PG = \left(\frac{\lambda}{4\pi}\right)^2 \frac{4}{R^4} \frac{(h_{BS} - h_m)^2}{|\chi|} \times \exp[2(H_T - h_m)k\,\mathrm{Im}\{\sqrt{\chi}\}]. \quad (18)$$

For glancing incidence it is again seen that the field varies as $1/R^4$.

At microwave and mm wave frequencies, attenuation through trees becomes quite large. As a result, paths involving diffraction around trees may give the most significant contributions to the received power. For example, propagation to an antenna in a forest clearing or road cut will take place via diffraction of the tree top fields down to ground level, rather than penetration through the trees. For mm waves, which have high diffraction loss, even isolated trees that block the propagation path can cause the failure of a communication link.

SEE ALSO THE FOLLOWING ARTICLES

IONOSPHERE • METEOROLOGY, DYNAMIC (TROPOSPHERE) • MICROWAVE COMMUNICATIONS • TERRESTRIAL ATMOSPHERIC ELECTRICITY • WAVE PHENOMENA • WIRELESS COMMUNICATIONS

BIBLIOGRAPHY

Balanis, C. A. (1997). "Antenna Theory, Analysis and Design," 2nd ed., Wiley, New York.

Bertoni, H. L. (2000). "Radio Propagation for Modern Wireless Systems," Prentice-Hall PTR, Englewood Cliffs, NJ.

Budden, K. G. (1985). "The Propagation of Radio Waves," Cambridge University Press.

Bullington, K. (1977). "Radio Propagation for Vehicular Communications," IEEE Trans. on Veh. Tech., Vol. VT-26, pp. 295–308.

Cheon, C., Bertoni, H. L., and Liang, G. (2000). "Monte Carlo Simulation of Delay and Angle Spread in Different Building Environments," IEEE Veh. Tech. Conference.

Crane, R. K. (1996). "Electromagnetic Wave Propagation Through Rain," Wiley, New York.

Devasirvatham, D. M. J. (1988). "Radio Propagation Studies in a Small City for Universal Portable Communications," Proc. of the IEEE Veh. Tech. Conference, pp. 100–104.

Du Castel, F. (1966). "Tropospheric Radiowave Propagation Beyond the Horizon," Pergamon Press, New York.

Felsen, L. B., and Marcuvitz, N. (1973). "Radiation and Scattering of Waves," Prentice-Hall, Englewood Cliffs, NJ.

Greenstein, L. J., Erceg, V., Yeh, Y. S., and Clark, M. V. (1997). "A New Path-Gain/Delay-Spread Propagation Model for Digital Cellular Channels," IEEE Trans. on Veh. Tech., Vol. VT-46, pp. 477–485.

Hall, M. P. M., and Barclay, L. W. (eds.). (1989). "Radiowave Propagation," Peter Peregrinus Ltd.

Hata, M. (1980). "Empirical Formula for Propagation Loss in Land Mobile Radio Service," IEEE Trans. on Veh. Tech., Vol. VT-29, pp. 317–325.

Janaswamy, R., and Andersen, J. B. (1998). "A Curvilinear Coordinate-Based Split-Step Parabolic Equation Method for Propagation Predictions Over Terrain," IEEE Trans. on Antennas. and Prop., Vol. AP-46, pp. 1089–1097.

Jordan, E. C., and Balmain, K. G. (1968). "Electromagnetic waves and Radiating Systems," 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Lecours, M., Chouinard, I. Y., Delisle, G. Y., and Roy, J. (1988). "Statistical Modeling of the Received Signal Envelope in a Mobile Radio Channel," IEEE Trans. on Veh. Tech., Vol. VT-37, pp. 204–212.

Liang, G., and Bertoni, H. L. (1998). "A New Approach to 3-D Ray Tracing for Propagation Prediction in Cities," IEEE Trans. on Antennas and Prop., Vol. AP-46, pp. 853–863.

Luebbers, R. J. (1984). "Finite Conductivity Uniform GTD Versus Knife Edge Diffraction in Prediction of Propagation Path Loss," IEEE Trans. on Antennas. and Prop., Vol. AP-32, pp. 70–76.

McNamara, D. A., Pistorius, C. W. I., and Malherbee, J. A. G. (1990). "Introduction to the Uniform Geometrical Theory of Diffraction," Artech House.

Mogensen, P. E., Eggers, P., Jensen, C., and Andersen, J. B. (1991). "Urban Area Radio Propagation Measurements at 955 and 1845 MHz for Small and Micro Cells," Proc. of GLOBECOM'91, Phoenix, AZ, pp. 1297–1302.

Okumura, Y., Ohmori, E., Kawano, T., and Fukuda, K. (1968). "Field Strength and Its Variability in VHF and UHF Land-Mobile Radio Service," Record of the Elect. Comm. Lab., Vol. 16, pp. 825–873.

Ott, G. D., and Plitkins, A. (1978). "Urban Path-Loss Characteristics at 820 MHz," IEEE Trans. Veh. Tech., Vol. VT-27, pp. 189–197.

Parsons, D. (1996). "The Mobile Radio Propagation Channel," pp. 134–135, Wiley, Chichester, England.

Pedersen, K. I., Mogensen, P. E., Fleury, B. H., Frederiksen, F., Olesen, K., and Larsen, S. L. (1997). "Analysis of Time, Azimuth and Doppler Dispersion in Outdoor Radio Channels," Proc. ACTS.

Sousa, E. S., Jovanovic, V. M., and Daigneault, C. (1994). "Delay Spread Measurements for the Digital Cellular Channel in Toronto," pp. 837–847, IEEE Trans. on Veh. Tech., Vol. VT-43.

Tamir, T. (1967). "On Radio-Wave Propagation in Forest Environments," IEEE Trans. on Antennas and Prop., Vol. AP-15, pp. 806–817.

Torrico, S. A. (1998). "Theoretical Modeling of Foliage Effects on Path Loss for Residential Environments," Dissertation, School of EE, The George Washington University.

Wait, J. R. (1996). "Electromagnetic Waves in Stratified Media," IEEE Press.



Radio Spectrum Utilization

Leon W. Couch, II
University of Florida

William J. Weisz
Motorola Inc.

I. Introduction
II. Radio Spectrum as a Natural Resource
III. Regulatory Organizations and Procedures
IV. Spectrum Allocation Principles
V. Use of the Radio Spectrum
VI. Users and Usage
VII. Usable Radio Spectrum and Radio Equipment
VIII. Technology and Usable Radio Spectrum
IX. Technology and Spectrum Utilization

GLOSSARY

Allocation of a frequency band Radio spectrum is divided into blocks of frequencies designated or allocated for use by a particular service or group of users such as public safety, cellular telephone, personal communication service (PCS), police, land mobile, television. Individual users are assigned the use of a narrow band of frequencies within the allocated block of frequencies.

Band Band is used to refer to a part of the radio spectrum; for example, the VHF band in the 150-MHz region, the cellular telephone band at 900 MHz, the PCS band at 1900 MHz, etc.

Bandwidth Bandwidth consists of the range of frequencies located between two designated frequencies. One of the frequencies is the lower edge of the band and the other frequency is the upper edge of the band.

Channels Radio channels are designated individual frequencies, or frequency bands that are assigned by a regulatory body. While television is allocated a block of frequencies between 174 and 216 MHz, there are seven television channels in this band; that is, channel 10 is assigned to occupy the frequencies between 192 to 198 MHz. Much of the radio spectrum has been assigned channel numbers; therefore, bands and channels are often synonymous.

Coverage Coverage is the geographical area in which the user can transmit and receive the desired signal at a level that provides adequate quality of the received information. For example, the coverage of a broadcast station is the area in which a user can receive a strong enough signal to provide a clear picture and sound. This can range from a few kilometers radius to several hundred kilometers radius. In some radio systems, coverage is extended by using several transmitters placed throughout the coverage area.

Radio frequency signal Signal on a specified designated frequency that transmits impulses or fluctuating electrical quantities, such as voltage, current, or electric

field strength, the variations of which represent coded information.

Spectrum allocation Process of assigning a certain part of the spectrum to be used for a particular radio service. It is common to refer to the entire set of assignments as the spectrum allocation.

Wavelength Speed of light divided by the radio frequency. A working approximation is 300 m divided by the radio frequency in MHz.

THE RADIO SPECTRUM is that portion of the electromagnetic spectrum that can effectively be electronically radiated from one point in space and received at another. Presently this includes frequencies from 9 kHz (kilohertz $= 10^3$ Hz) to 400 GHz (gigahertz $= 10^9$ Hz). While there are international allocations for the radio spectrum up to 275 GHz, most of the commercial use takes place between 100 kHz and 20 GHz and very few experimental systems have been operated above 100 GHz. Signals using these same frequencies can also be generated and conducted via wires, coaxial cables, and glass fibers extending from transmitter to receiver, but since such signals are contained and not intended to be radiated, they are not part of the allocated radio spectrum. There are many sources that unintentionally radiate, for example, the sun, power lines, computers, and automobile engines, and these must be considered because they impact radio spectrum usage. Since the same radio spectrum is shared by all within a given coverage area, it constitutes a common "natural resource." Allocations of portions of the radio spectrum are set worldwide by periodic international conferences and administered by the International Telecommunications Union (ITU), between the various national telecommunications authorities, who determine national policy for each country. Spectrum usage, or allocation, is determined by type of use, or "service," and is managed so as to optimize usage.

I. INTRODUCTION

Radio spectrum usage includes all of the transmissions made and signals that are received. The utilization of the radio spectrum is a measure of how effectively it is used. Efficient utilization of the radio spectrum depends on optimizing its use, but because of the various types of services using the spectrum, there is no universally accepted single method of measuring utilization and comparing the efficiency of one type of use versus another.

In the broadcast service, high utilization comes from providing as many broadcast channels as possible in a given area without causing destructive interference to the listening public. This is done by establishing frequencies or channels and controlling characteristics of the broadcast equipment used. Each channel has only one transmitter operating in any given coverage area, and many receivers.

In services such as Land Mobile (police, fire, business use, etc.) and cellular telephone, high utilization comes from providing service to a large number of roving and fixed users in a given area, while providing a coverage over the working area that will give the quality of service required without causing destructive interference between users.

In some services, 100% receipt of signals is mandatory. Normal statistical methods of measuring spectrum utilization do not apply. In these instances, utilization may be very high despite the fact that there are very few transmissions. The distress calling channels used by ships and airplanes are examples of such cases. Here, the emphasis is on keeping all unnecessary transmissions off these channels so that the distress calls can be received, even under adverse conditions.

II. RADIO SPECTRUM AS A NATURAL RESOURCE

Radio spectrum is universally considered a natural resource. In the industrialized areas of the world this resource is extensively used and there is constant pressure to regulate radio spectrum to make room for additional users. In contrast, in many developing countries, and in areas of low population density, portions of the radio spectrum are virtually unused.

Nevertheless, radio spectrum usage is increasing everywhere. Not only are more users seeking to use existing services, but also there are new services that require portions of the radio spectrum. These new and different services will place future demands for radio spectrum greater than those previously envisioned.

The radio spectrum is a special type of natural resource that can instantly be renewed. It is not permanently depleted like oil. If everyone using the radio spectrum were to stop transmitting, the entire spectrum would be instantly available for other uses. This has only happened, however, during wartime, or when a regulatory body changes a group of users from one part of the radio spectrum to another.

On the other hand, the radio spectrum resource, or parts of it, can be effectively depleted if the number of transmissions exceed the capacity of the band. When this happens, mutual interference becomes so great that the amount of usable transmission is reduced below the point of usefulness. Where this has occurred, the cause can be traced to the national regulatory body having permitted too many users access to a band, or the lack of either clear operating

rules or operator discipline. For this reason, national governments grant users the privilege of using radio spectrum in exchange for agreements to abide by usage rules. In addition, the users often agree among themselves on informal additional procedures to ensure that the spectrum resource is not overloaded.

In time of disaster, man-made or natural, the emergency demand for additional transmissions rises to several times normal. At these times even the best disciplined services may overload if radio frequency allocations have not been carefully made to accommodate increased usage needs. The regulatory body must always keep this potentially large demand in mind when considering allocations. The radio spectrum resource reaches its maximum value during and just after such an event.

As with many natural resources, the radio spectrum resource has no real value until it has been developed through generally interference-free use, that is, until equipment for the specified use has been designed, built, and placed in operation. Radio spectrum then acquires the value of the equipment and the service that it provides. As a result, radio spectrum allocations, once made, tend to be extremely resistant to change.

The potential value of a portion of the radio spectrum is a function of the ability to use it economically and efficiently. This depends on many things such as frequency band service, and the locale etc. The market value of a single television channel in a major city is in the hundreds of millions of dollars. The worldwide capital investment of the public in over 480 million television sets and broadcast equipment of thousands of TV stations provides great incentive to retain the current allocation for television. The same can be said, to some degree, for every established service in the world. Where the radio spectrum is used by commercial businesses as a tool to increase productivity, cost saving can generally be accurately determined.

In contrast, spectrum used to transmit timely and accurate information to and from policemen, firemen, and soldiers may have a value not measurable in direct monetary terms. However, a large investment must nonetheless be made in equipment and training to provide for this capability. This investment can be lost if the service is required to change to different frequencies or transmission types.

National governments undertake the task of evaluating the needs of government services (either local or national), together with the needs of commercial services, in order to appropriately allocate and enforce specific radio spectrum uses. Their task is complicated by requests for spectrum for new services resulting from technology advances as well as maximization of return on the existing users' investments. Most countries have formed an agency (in the United States, this is the Federal Communications Commission) to administer the radio spectrum. However, some countries rely on a state communications company to allocate the radio spectrum. Other countries give the responsibility to the military.

In the lower frequency bands, for example, below 20 MHz, and in certain higher frequency bands, such as those used for satellites, the signals can at times reach around the world across national borders. The maritime and aeronautical use of radio spectrum requires that the transmitters aboard airplanes and ships, moving from one country to another, must operate under the jurisdiction of the country in which they are operating. It is, therefore, necessary that there be worldwide agreement on the frequency allocations in many bands. Because radio frequency signals often cross national borders where they could interfere with radio frequencies allocated in another country, nations have additionally had to find ways to coordinate their individual allocation procedures. The International Telecommunication Union (ITU), an organization of the United Nations, has been formed for this purpose. The ITU seeks and obtains agreement on the broad categories of services to operate in specified radio frequency bands. The nations belonging to the ITU accept these agreements as treaties. The failure of governments to abide by international allocations, or to enforce these agreements, could destroy the value of the service using that portion of the radio spectrum.

III. REGULATORY ORGANIZATIONS AND PROCEDURES

A. International Regulatory Organizations

The International Telecommunications Union (ITU), part of the United Nations, is the international body that determines the worldwide radio spectrum allocations. It does this through World Administrative Radio Conferences, where ITU member nations participate in the decision making. The consensus allocations developed at these conferences are ratified by member nations by treaty. The resultant tables, published by the ITU, are known as the Radio Regulations.

The Radio Regulations indicate which service is to use a given frequency band. For some bands, the rules indicate that two or more services may share a given band. Each country then makes its own detailed frequency allocation plan consistent with The Radio Regulations tables.

Many fixed service systems are intended to transmit information internationally. This requires that the equipment be mutually compatible. The Radiocommunication Sector (ITU-R), a part of the ITU, is organized to study technical and operating questions in radio communications and to issue recommendations for equipment and signal specifications. Among such recommendations are those

on transmission specifications that will provide the required intercountry compatibility. This committee also provides recommendations for radio system technical requirements that are used by many countries to help them in planning their own internal allocation of the radio spectrum.

Many services use a channel designation to identify the authorized transmitter frequency. For example, television stations are known to operate on channels rather than frequencies. The ITU-R makes recommendations for channel frequency spacing and numbering plans, which, when adopted, aid aircraft and ships that may have to operate worldwide.

History shows that once a service has been established and uses a particular portion of the spectrum it is seldom changed. When a new service or use for the spectrum is developed, the regulatory agency often assigns it to high frequencies. This is the most prevalent reason why an examination of the allocation table shows that many services have allocations widely scattered over the spectrum.

B. National Regulating Agencies

In the United States, the regulation of radio spectrum is done by two agencies of the federal government. They are responsible not only for spectrum allocation, but also for any technical equipment specification necessary. The Federal Communications Commission (FCC) regulates radio spectrum for all non-federal government use, including the aeronautical and maritime mobile services. The National Telecommunications and Information Administration (NTIA) regulates spectrum used by the agencies of the federal government, including the military.

The NTIA does not have to make public the reasons for its allocation of spectrum. Its task is complicated by the necessity of coordinating with all countries in which the United States has received permission to operate governmental agencies using radio spectrum (e.g., embassies and military bases).

On the other hand, the decisions of the FCC are based on public participation where any interested party can make proposals and comment on any and all proceedings. Because the radio spectrum is a valuable resource, there are many vested interests, both personal and institutional, that impinge on any allocation changes. The U.S. system, which has by and large worked well, is by its very nature a political process. After receiving all of the proposals and comments, both written and oral, the FCC commissioners must often make difficult allocation decisions. These decisions are subject to judicial review. FCC actions fall under the *Telecommunications Act of 1934*, and all subsequent Federal laws. One advantage of the U.S. system is that the public's needs are constantly reviewed and are a part of every decision. An advocate for change can advance his cause at every level of the process. One disadvantage is that the system may not be as quick to respond to technology changes or to new services.

As a matter of policy, the FCC has not regulated receiving equipment performance specifications to the same degree as regulatory organizations in other countries. It does, however, regulate transmissions that directly affect spectrum usage. This is because a receiver owner can choose his own degree of communication selectivity and quality without causing interference to other users or affecting spectrum utilization.

In many European countries, the responsibility for radio spectrum allocation is given to state communications monopolies that may also operate national telephone systems. Some of these organizations operate on the principle that they have the responsibility and the right to provide any and all of the communication needs of their citizens. Therefore, they allocate the radio spectrum in ways that best allow them to provide services as they see fit.

In general, these organizations have exercised strong system engineering influences, and regulate more aspects of the user's radio system, including both receiver and transmitter performance specifications.

Every year there are new requests for services that require radio spectrum as well as an increased demand for the existing services. While these demands can easily be met in less developed parts of the world, it is a real problem in metropolitan areas of developed countries. A great deal of pressure is placed on regulatory agencies in developed countries to find new and better ways of allocating or reallocating the sought after radio spectrum resource.

C. Allocation Procedures

Frequently, more people seek to be licensed to use a frequency band than there are frequency allocations available. In many countries, only government agencies can provide the communications services. In these cases there is no problem in deciding who should be licensed to use radio frequencies. However, more countries are putting the broadcast and communications business in the hands of private companies. Therefore a government-issued license to use a frequency or band of frequencies is necessary for an organization that wants to operate a radio system, be it for public safety, business, or other use.

In the United States, the Federal Government encourages the development of many radio services by individuals or companies. As a result, broadcast licenses, for example, are transferred (with the FCC's approval) when a broadcast station is sold. This license is a tangible asset of great value. When new allocations are available, the FCC first determines which applicants are capable of providing

the service. Where there are more qualified applicants than there are available allocations, the FCC can license those it feels are best capable of providing the service. Financial capability is a major factor when the service requires a big investment in equipment and personnel.

In recent years, there have been more highly qualified applicants than allocatable frequencies. In an attempt to be fair, the FCC has turned to the use of lotteries to choose among qualified applicants. Alternately, for selection of PCS providers, the FCC conducted auctions and sold the licenses to the highest bidder.

While this allocation technique might determine the value of the natural resource and shift the usage to that part of the spectrum that could produce the highest dollar return (some people believe this is maximum spectrum utilization), it may not provide equal access to the spectrum by those who need it the most and are not able to acquire the necessary financial base to compete for the use of the spectrum.

Many other countries face the same problems and could adopt this processes or develop other ways to extract revenue from the use of radio spectrum.

IV. SPECTRUM ALLOCATION PRINCIPLES

In most countries, the priority followed by the regulators for allocating radio spectrum is as follows.

1. Military
2. Aeronautical and maritime emergency communications, radio navigation, and radio location
3. Public safety functions, for example, police, fire, and other emergency services
4. Telecommunication companies, for example, cellular and PCS systems, microwave and satellite telephone relays
5. Broadcast radio and television
6. Private users, for example, private mobile systems and private fixed microwave relays
7. Others.

For locations where there is high spectrum usage, regulatory agencies must determine if services can be provided using means other than radio, such as fiber optic cable, coaxial cable, or twisted pair lines. They must also consider the relative cost of using alternative methods of providing the service. For example, many studies have been conducted to determine the feasibility of providing radio and television broadcast service by coax or fiber cable. Such a change would allow the regulators to reallocate much valuable spectrum to other services. Such changes, however, must also consider the cost of scrapping existing equipment. While public necessity is the first priority (see earlier), the public interest and convenience is a factor in many services such as broadcasting, cellular telephone, the amateur and citizen radio services, and microwave ovens. While these services are not necessary to save lives, they do provide for an improved quality of life and are desired by the public.

V. USE OF THE RADIO SPECTRUM

For regulatory purposes, uses of the spectrum are classified by the ITU by type of service provided. The major services and some subclassifications are as follows.

A. Broadcast Service

The radio transmissions in this service are intended for direct reception by the general public and generally are radiated from one antenna location to cover a wide area around the transmitter where thousands of randomly located fixed or moving receivers operate. The familiar AM and FM radio, as well as television transmissions, are in this service.

B. Fixed Service

This service constitutes radio communications between fixed points. The most common equipment used in this service is telephone microwave systems used to interconnect switching centers.

C. Mobile Service

The radio communications in this service are between and among automobiles, trains, aircraft, person-carried hand-held units, etc. and fixed stations.

1. Aeronautical Mobile Service: Radio transmissions to and from airplanes and including emergency radio beacons
2. Land Mobile Service: Business and Public Safety radio transmissions to and from land vehicles and portable units
3. Maritime Mobile Service: Radio transmissions to and from ships
4. Cellular Telephone and PCS: Personal radio communications from vehicles and hand-held units to base stations connected to the Public Switched Telephone Network (PSTN).

D. Satellite Service

The broadcast, fixed, and mobile services have corresponding satellite services. In addition, there are satellite

services for amateur, meteorological, standard frequency, time, earth exploration, and radio navigation. The transmissions in this service generally start from earth stations and are relayed by a satellite station back to earth. In some uses the satellite may originate communication.

E. Standard Frequency and Time Signal Service

A radio communication service for scientific, technical, and other purposes, providing transmission of specified frequencies, time signals, or both, of stated high precision, intended for general reception.

F. Radio Navigation Service

Transmissions in this service are for navigation and obstruction warning.

G. Radio Location Service

In this service the radio transmissions are used to determine position, velocity, and other characteristics of objects. Radars, radio altimeters, and the Global Positioning System (GPS) are examples.

H. Radio Astronomy Service

The transmitters in this service are of cosmic origin such as stars. Because the signals received have very low power, many regulators have provisions in their allocations to protect certain radio frequency bands in locations that are near radio telescope sites. They do this by prohibiting or restricting the power of transmitters that could radiate power in the protected radio frequency bands.

I. Amateur Service

The amateur service is for self-training, intercommunication, and technical investigations carried out by duly authorized persons interested in radio technique solely with a personal aim and without pecuniary interest. Almost all countries have recognized that by encouraging radio amateurs they create a group of people trained in the use, maintenance, and design of radio equipment. Amateurs have consistently shown the practical use of the higher radio frequencies before there were commercial applications at these frequencies.

J. Industrial, Scientific, and Medical (ISM)

Many machines use radio frequency power for purposes other than transmission of information. Such equipment is used to generate heat, control chemical reactions, and make measurements. Microwave ovens are one example. The very nature of ISM equipment mandates that it radiate large amounts of radio frequency power. While there are no regional allocations for this service, the International Radio Regulations have specified that ISM equipment radio frequencies be restricted to certain bands.

VI. USERS AND USAGE

A. Broadcast Service

Broadcasting is the most widely used radio service. Its allocations include a variety of frequency bands with various types of transmission capabilities. While licenses to operate broadcast transmitters are granted to individuals or institutions (see Table I), the real beneficiary is the general public. The oldest broadcast service uses amplitude modulated signals in a frequency band that extends from 535 to 1705 kHz (the AM broadcast band). At certain times, such AM broadcast signals can propagate unusually large distances, especially after sunset. Such signals can cause destructive interference in broadcast receivers tuned to receive other desired weak signals. International cooperation in frequency allocation, regulating power and determining the daily hours of broadcast has been extensive, and service to the public has been excellent.

FM radio and television are also widely used for broadcast. These systems use frequencies above 54 MHz and are much less susceptible to the long distance "skip" interference characteristic of AM broadcast above. In order to provide the users with high quality signals, stations with overlapping coverage areas are not assigned the same frequencies.

FM broadcast is resistant to interference to strong signals by weaker signals, so the same channels can be reassigned at distances of 150 km or more separation between stations.

B. Fixed Service

The fixed service is least known to the general public, but is an essential part of the radio industry. Systems in the fixed service are, for the most part, operated by communications companies or by companies that have a need for large

TABLE I Stations Authorized by the FCC for the Year 1998

Service	Number of stations
Broadcast	
AM	4793
FM	7679
TV	1589

communications networks—for telephone or data communications. Since these systems transmit information between fixed points, there is a constant reevaluation to determine whether coaxial cables or glass fibers can more effectively be used. In the urban areas of the industrialized countries, most fixed services use microwave frequency bands above 5 GHz, 5 to 20 MHz in width. Relay stations are used for long distance microwave communications to receive and retransmit information. These relay stations are spread 40 to 50 km apart along a line between the end points of the network. Because of the wide bandwidth, microwave systems have large capacity and are capable of simultaneously transmitting several television signals and thousands of telephone messages. Relay stations can cost well over a million dollars and the cost of the equipment at each end can cost tens of millions of dollars.

In less developed countries and areas of low population density, equipment using lower frequencies (150–500 MHz) and narrower bandwidths (25–100 kHz) are also used for the fixed service. In these areas the fixed service is used to extend telephone services beyond the wire network into rural areas.

Frequencies below 20 MHz are also used in the fixed service to provide both telephone and data service to isolated areas such as remote islands.

There has been a long history of regulatory changes that move the fixed service out of the mid-frequency bands (50 MHz to 2.5 GHz), and to allocate these frequencies to the mobile services. Today, more of the functions that are provided in the fixed services are being transferred to the satellite services which are more reliable and often less expensive.

C. Mobile Service

The mobile service consists of three different services: aeronautical, mobile, and maritime. While there are possible alternatives to radio for providing the fixed and broadcast requirements, radio is the only means of providing communication to units or people in motion. The need to communicate with moving vehicles and people has made the mobile service the fastest growing segment of all users of spectrum.

While these services have dissimilar needs, they have the following characteristics in common:

1. They provide communication between and among land-based, fixed locations and airplanes, vehicles, ships, or people.
2. They began as a means of providing emergency communication, and still retain that as a first priority.
3. The communications requirements of this service have expanded until today most of the information transmitted is for personal convenience as well as for administrative and operational purposes. The business users of this service are constantly finding ways to use this transmitted information to make their mobile fleets more efficient and effective. Fleets that have used radio extensively have shown significant improvements in productivity for both personnel and equipment. The use of radio is now necessary to be competitive.
4. Their transmissions are mostly of short duration and any one user may only need to transmit for a few seconds at a time, adding up to only a few minutes per hour. This is in contrast to the broadcast and fixed services where transmitters are on continually or for long periods of time. This intermittent use characteristic allows many users to share the same frequency band in the same location.
5. Cellular systems are often used where subscribers communicate through a base station which is connected to other base stations. Each base station covers a small geographical area called a cell. Frequencies are reused in adjacent cells to accommodate a large number of subscribers. As the subscriber population increases, the cell size can be made smaller and more cells added to accommodate more subscribers.

An increasing amount of frequency usage in the mobile service is for transmissions between people in motion carrying small hand-held radio equipment. The need for this type of communication in all of the mobile services is growing at a rapid rate, especially in urban areas.

The aeronautical mobile service is used to communicate with airplanes. The familiar ground control communications with aircraft is allocated frequencies between 108 and 144 MHz. Lower frequencies (between 4 and 6 MHz) are used to maintain contact with transoceanic flights as they move out over the oceans.

Cellular telephone services are the most widely used land mobile applications in many countries. This extension of the telephone service uses many frequencies between 400 MHz and 2.5 GHz. In the United States, one variation of this service, called the Cellular Telephone System, uses frequencies in the 900-MHz band and PCS uses the 1900-MHz band. These services provide high quality mobile telephone service to millions of private and business people and adds greatly to radio spectrum requirements of the mobile service.

The demand for inexpensive equipment to communicate to and from the automobile has resulted in small spectrum allocations for "citizens' services." In North America, the 27-MHz citizens' band service and the 460-MHz Family Radio Service (FRS) have been established where one can operate low power equipment without formal licensing. A great many radios have been sold for these services. In Japan, a similar "personal radio service"

at a frequency near 900 MHz has attracted thousands of users.

Personal paging is a one-way land mobile service. The paging service users wear small receivers and the system provides for the transmission of a message to the user from any telephone. The message may be a short voice message or a message that can be read from a visual display. These systems operate on frequencies up to 1 GHz.

The maritime mobile service has allocations throughout the radio spectrum. This service has two types of operations. When the ships are in or near ports, they communicate on frequencies above 146 MHz. At sea, they use frequencies below 26 MHz. More than in the other mobile services, most of the information transmitted to ships is data. Radio teletype and other automatic data systems are used. Although the speed of these systems is slow, it is well adapted to the propagation characteristics of the low frequencies used. A large number of ships now use satellite relays for transoceanic communication. These satellite systems are more reliable and allow for higher rates of information transmission.

D. Satellite Service

Satellite systems are in service, or have been proposed, for each of the other services. Most satellite usage, however, is for the fixed services where satellite relay systems are operated by international or national communications companies. These systems transmit much the same information that is transmitted by fixed service microwave systems (e.g., television programs and telephone messages). Some satellite systems now operating provide telephone service to outlying regions of less developed countries as well as all over the world.

There are also national systems in the Satellite Broadcast Service that provide television broadcast directly to homes. Many specialized broadcast features, including movie channels, are available to those willing to incur the cost of program fees. They use inexpensive receivers and small dish antennas (1/2 in diameter).

Most of the fixed, broadcast, and maritime mobile services use synchronous orbits and frequencies above 1.5 GHz.

E. Radio Navigation Service

Radio navigation services serve aeronautical and maritime users. Radio navigation systems consist of networks of transmitters that allow the user to locate his position in relation to the network, and hence themselves. The Global Positioning System (GPS) consists of a network of low earth orbit satellites. It is one of the most popular systems since inexpensive hand-held GPS receivers are readily available.

In navigational systems the transmitters broadcast sufficient information so that the location can automatically be calculated and displayed. In addition, there are specialized beacons and location markers, like those used at airports, to provide guidance for an aircraft.

Most navigational systems allocations range from 9 kHz to 2 GHz. GPS navigation systems are increasingly used in many applications.

F. Radio Location Service

Most frequency usage in this service is for radar. Since radar obtains its directional information from the use of very narrow antenna patterns, and since such narrow antenna patterns are more easily achieved at higher frequencies, most of the radars operate at microwave frequencies above 2 GHz. Radar obtains its distance information by measuring the time for a radio pulse to travel to and from the target. To improve the accuracy of this measurement, the radio pulses are made very short, thus increasing the bandwidth occupied by the radar transmitter. The bandwidth can be several MHz wide.

G. Amateur Service

Radio amateurs (or "hams") were the first users of the radio spectrum and originally had use of the entire spectrum. The amateur service is the only service that had continuously given up spectrum to other services. As amateurs demonstrated the practical use of the higher frequencies, they were found to have commercial value and reallocated to new uses. While much amateur spectrum has been transferred, there are still allocations for amateur bands that extend from 1.8 MHz to 250 GHz. Most countries certify a user by means of a technical examination before he/she is allowed to use the spectrum. Over the years, the amateur radio operator has demonstrated a high degree of technical and operational discipline.

VII. USABLE RADIO SPECTRUM AND RADIO EQUIPMENT

To understand the factors that must be considered in order to make optimum use of the radio spectrum, it is necessary to understand a little about spectrum and radio equipment characteristics that impinge on such judgments. Because spectrum allocations are often affected by the cost of equipment to produce and receive transmissions, some brief comments on such equipment will also be made. The most basic characteristic of the radio spectrum is the location of a frequency in the spectrum. There are many factors to be considered that are a function of frequency

TABLE II Radio Characteristics versus Frequency

Characteristic	Lower frequency	Higher frequency
Amount of radio frequency noise	Greater	Less
Ease of generating radio frequency power	Easier	Harder
Ease of radiating radio frequency energy	Harder	Easier
Ease of directing radiated radio frequency energy	Harder	Easier
Propagation loss of radio frequency energy over land or sea	Lower	Higher
Multipath degradation of the received signal	Less	Greater
Ease of rejecting unwanted signals	Easier	Harder
Ease of keeping transmitter or receiver on assigned frequency	Easier	Harder

when determining the use that will be made of a specific band of frequencies. In general, the relationship with frequency that exists is shown in Table II.

A. Radio Frequency Noise

Noise is present at all frequencies in the radio spectrum and consists of signals having random amplitude and phase fluctuations. Sources of noise include the sun, the stars, the atmosphere, for example, lightning, and man-made equipment, including radios. Both naturally generated and man-made noise is greatest at the low frequencies and decreases at the higher frequencies, and man-made noise is greater in urban than in rural areas. Below about 1 GHz, external noise is the limiting factor in the effective sensitivity of a receiver.

The amount of noise a receiver is exposed to is proportional to the bandwidth of the receiver. While a well-designed receiver having a bandwidth that is typically found in the land mobile service requires a signal only slightly greater than the noise, an analog television receiver requires a signal that is much larger than the noise.

B. Radio Transmitter Power

The cost of radio transmitters increases with both frequency and power. At the low frequencies, most transmitters have a low power, stable, accurate frequency source that may be modulated with the intelligence to be transmitted. This signal's power is amplified 1,000,000 times before it is applied to the antenna. The transmitter radio frequency output power can range from a few hundred milliwatts to hundreds of thousands of watts. At the low frequencies, amplifiers that produce a few watts are not generally expensive, but higher power amplifiers are costly. At higher frequencies, the cost of amplifiers increases and output radio frequency power decreases. At frequencies above 2 GHz, it is very expensive to use amplifiers that produce more than a few watts.

While it is possible to configure a transmitter to generate radio frequency power in a single device to be useful for community information, the frequency of the device must be controlled accurately and the device modulated by the intended intelligence. This takes additional, often expensive, circuitry.

One well known noncommunication appliance, the microwave oven, uses the single device method and is able to generate hundreds of watts. However, the oven frequency is not modulated and its frequency is allowed to vary over a considerable band of frequencies.

C. Radiating Radio Frequency Energy

In order for the radio frequency generated by a transmitter to be useful, it must be transformed from electrical energy in the transmitter to electromagnetic fields in space. This is the function of the antenna. Effective antennas are at least one-fourth to one-half of the wavelength of the frequency being used. Antennas can be improved by making them larger, but the size of such antenna structure may be impractical for long wavelengths. Effective antennas for AM broadcast stations may have a height of over 200 m. Some antenna structures for very low frequencies (e.g., 200 kHz) have lengths of over 1 km. Geographical coverage range in the higher frequencies can generally be extended by mounting the antenna on a tall tower or on the roof of a building.

At the high frequencies (above 2 GHz) the antenna dimensions become very small and difficult to work with. At these frequencies a properly flared opening (horn-type antennas at the end of the transmission line) can provide effective radiation of these frequencies.

Generally, it is easier to radiate radio energy at higher frequencies because of the small size of the antenna structures required.

While the above statements have been directed at transmitting antennas, they apply equally to receiving antennas as well.

D. Directing Radiated Radio Frequency Energy

In addition to radiating radio energy, antennas have directional characteristics. The directivity of an antenna is a very important system parameter. For many services, antennas should radiate equally in all directions of the compass with as much of the energy directed at the horizon as possible. Broadcast, maritime, and land mobile stations are a few examples of such services. In such cases, the coverage area is a circle centered at the antenna.

On the other hand, most stations in the fixed service transmit almost all their energy in a narrow beam only

in the direction of the intended receiving station. There are several system benefits that result from this: all the transmitter's power is focused in one direction and therefore less power is required to deliver a usable signal level to a receiver. At the receiver, interference is reduced if it originates from a different direction than the intended transmitter. Because the transmitter's power is focused in one direction, this allows the same radio frequency to be used in the same geographical area several times by other unrelated fixed stations whose transmitter power is focused in other directions. The impact on the allocation of frequencies depends on the degree of directivity that can be obtained.

The degree of directivity that an antenna can have is a function of the dimensions of the antenna. The dimensions of directional antennas range from 1 to 1000 wavelengths. Television receiving antennas are about two wavelengths long. Fixed station antennas for microwave stations can be up to 1000 wavelengths across. While it is possible to build directional antennas for the very low frequencies, the antenna structures become very large, possibly up to several kilometers long.

For some services, it is desirable to point the directional antenna in different directions at different times. Examples of this are amateur stations and television receivers. Radars in the radio location service use antennas that are highly directional, but are constantly rotated. Most satellite, earth, and space stations use directional antennas and many are steerable.

Whenever an antenna has directivity, it almost always has an effective gain. This is measured as the effective field strength radiated by a directional antenna in a given direction, compared to the field strength produced by an omnidirectional antenna that radiates the same power equally well in all directions. Typical gains range from 2, for some television antennas, to over 1000 for some fixed station antennas.

It is easier to obtain high directivity at the higher frequencies because of the shorter wavelengths and hence smaller antenna dimensions required.

E. Propagation Loss of Radio Frequency Energy over Land or Sea

At the low frequencies (below 500 kHz), radio frequency energy propagates through earth, sea, and the atmosphere. At these frequencies, propagation loss over sea is about three times less than that over land and communication circuits of over 1500 km are readily established. This has made these frequencies very valuable to the maritime service for communication to ships at sea and to the fixed service for communication to remote islands.

At higher frequencies (2–20 MHz) the land and sea propagation losses increase at a rate faster than the atmospheric losses. At times, longer communication circuits can be established using atmospheric radio wave reflections, but the loss varies with time of day and such communication circuits cannot be reliably maintained for long periods of time.

Although the loss at one frequency may increase, the loss at another frequency may decrease. The allocation of these frequencies is made in such a way that communication systems can switch from one band to another to find a usable frequency. Because these frequencies are so very useful, many international agreements exist for the allocation of these frequencies.

It is of interest to note that the information bandwidth of a single satellite relay station, which operates at a frequency around 20 GHz, is about as great as the entire spectrum from 9 kHz to 20 MHz and such a satellite can provide communication service to ships and remote islands more reliably.

In the frequency range between 30 MHz and 1 GHz the radio propagation approaches that of light. At 30 MHz there is enough refraction of the radio waves that transmissions are possible beyond the horizon by as much as 100%. As frequency increases, transmission becomes more line of sight and is limited more and more to the horizon. Increased propagation loss is also incurred in places that are shadowed from the transmitter by hills or buildings. For frequencies greater than 1 to 2 GHz the propagation loss is very great between locations that are not line of sight.

F. Multipath Degradation of the Received Signal

At low frequencies (500 kHz) the energy propagated through the earth and the energy propagated through the atmosphere travel different distances and can arrive at the receiving antenna out of phase, and, to some degree, can cancel one another. This can have the effect of increasing the propagation loss. If the antenna is moving, the cancellation will slowly change to reinforcement and back to cancellation. The distance between cancellation may be several kilometers. While the effect is very noticeable, the slow rate of change minimizes the degradation to communications.

At higher frequencies (e.g., 50 to 1000 MHz), the antenna may receive signals from many reflectors, hills, buildings, overhead wire—which can combine and cause cancellation, and the effect is the same as for the low frequencies, except that the cancellations occur every 6 to 0.3 m. The effect is most noticeable when the magnitude of the direct signal is about the same as the reflected signals.

For the land mobile service, this can be the case over much of the coverage area. If the antenna is situated on a fast moving vehicle, it can pass through many cancellations in a second, disrupting the communication link. The degree of degradation of the received signal increases with the frequency of the signal and the speed of the vehicle.

To overcome this degradation, systems are designed to have a higher field strength in the coverage area. This affects the distance between systems before the frequency can be used again and must be accounted for in the spectrum allocation process.

G. Rejecting Unwanted Signals

Every receiver has some capability to select a desired signal and reject all others. This is called receiver selectivity. The better the selectivity, the greater the receiver cost. If, in a given coverage area, AM broadcast transmitters are assigned to adjacent broadcast channels, receivers would have to have a high degree of selectivity in order to clearly receive a remote station if the receiver was physically close to the adjacent channel broadcast station. In practice, this is never the case, since spectrum allocation regulations regarding geographic spacing of adjacent channels assure that in any location there are no broadcast stations assigned closer than four or five channels to each other. This reduces the required selectivity of broadcast receivers, with correspondingly lower cost, at the expense of using additional spectrum. A good rule of thumb is that better spectrum utilization in any given frequency band generally requires costlier equipment.

The frequency band between the assigned channels is called the "guard band," and almost all allocations have provision for some amount of guard band. The guard bands can be reduced if transmitters are not physically closely spaced. This, however, is not practical in the mobile services where the mobile transmitters may transmit from any location. While it is not possible to build receivers with perfect selectivity, professional receivers are available which can reject unwanted interference signals only 25 kHz away from the desired signal and which are 100,000 to 1,000,000 times greater than the desired signal.

VIII. TECHNOLOGY AND USABLE RADIO SPECTRUM

Commercial use of spectrum is limited by the cost of generating the frequency being used with sufficient power and of receiving it. Almost without exception, the cost of devices to achieve transmitter power increases with the frequency and the power level required. For these reasons, the first commercial application of a higher frequency is usually for system applications that can more readily support the use of the costly devices. When the technology progresses and the cost of the devices is reduced, other system applications for that frequency are found. For example, frequencies between 500 and 1000 MHz were first used for the fixed services where the cost of the devices could be spread across many telephone channels. The devices used were low power—less than a watt. When the technology allowed one to build costly high power devices (greater than 20,000 W) at these frequencies, television transmitters were built. When the technology advanced further so that solid state devices of moderate power (up to 50 W) were available, they found use in the private land mobile service. Now, 0.6-W hand-held cellular phones that operate at 900 MHz are cost effective.

IX. TECHNOLOGY AND SPECTRUM UTILIZATION

A. Higher Frequency Practicality—An Example: The Land Mobile and Cellular Phones

The land mobile service began as a police radio service using compatible modulation at 1600 kHz. In 1940, FM systems were permitted and began operating in the 30 MHz area. In its early years the police band was allocated on the basis of a scaled down version of the FM broadcast system. The FM system proved to be superior and soon was the most popular two-way police radio system. This popularity, and subsequent advances in technology, resulted in several new allocations for police use higher in the radio frequency spectrum. In 1985, the United States, channels are allocated for police use in the frequency bands of 30 to 50, 150 to 170, 450 to 512, and 806 to 866 MHz. Now, with advances in technology, low-cost cellular phones operate at 900 MHz and PCS phones operate at 1900 MHz.

B. Digital Signals

Until a few years ago, almost all the services used analog modulation to transmit information. One characteristic of this type of transmissions is that every time the signal is processed or relayed, some noise and distortion is added. While this can be minimized, it cannot be eliminated. Some signals may be processed 30 to 40 times as they move from source to destination. Digitizing, wherein the value of the signal is represented by a series of symbols, usually on or off pulses, allows systems to be designed that eliminate reprocessing errors. While the process of digitizing signals has many advantages in wire, coaxial cable, and radio systems, it often requires the use of a greater bandwidth in order to utilize the advantages.

The trend in digital systems is to quantize the signal at its source and not to decode the signal until it is at the location where it will be used. The digitized signal can be transmitted over many media such as wireline, microwave, and satellite relays.

Because digital signals offer the user higher quality transmissions and greater freedom in processing the signals, many of the new services are now digital. In addition, there are trends to convert existing systems to digital modulation. Some countries are allocating large bands of frequencies to be used by systems that will use only digital signals and will provide both fixed and mobile services. While digital systems generally use a greater bandwidth than analog systems, the digital systems tend to be less degraded by low level unwanted signals. Also, new technology allows multiple digital signals to share the same frequency band. Here, a wideband coding signal is used to spread the spectrum and multiple signals (users) can share the same band by using orthogonal codes. This is called spread spectrum (SS) or Code Division Multiple Access (CDMA). This allows digital systems to achieve a utilization greater than the analog systems.

C. Additional Information in Allocated Bandwidth

Many radio systems, when first developed, use more bandwidth than necessary due to technological limitations or equipment cost limitations. These systems are given allocations based on the original equipment's bandwidth needs. As stated before, once an allocation is made, it is difficult to change. On the other hand, as technology advances, it may be possible for additional usage to be made of the same bandwidth. Well-known examples of this include the addition of color to television, the addition of stereo to FM broadcasting, and the addition of stereo to AM and television broadcasting.

Newer examples are digital television and digital radio broadcasting. Using data compression in digital TV broadcasting, one high definition picture or five standard definition pictures can be transmitted over the same 6-MHz-channel bandwidth that was previously assigned to one analog TV station. Digital audio broadcasting allows CD quality audio as well as additional textural information (such as the station call letters and the title of the program) to be transmitted.

D. Allocation Alternatives To Accommodate Growth and New Users

Many radio services need to accommodate the growth of new users. Also, enhanced service such as data transmission in the land mobile service needs to be accommodated.

1. How Growth Has Been Handled—An Example

The cellular phone service is a good example of how new users can be accommodated. A single cell site serves a small geographical area and a large number of cell sites allow coverage of the whole area. The frequencies used in one cell site can be reused in nearby cell sites to accommodate additional users. As the user density increases, the size of the cell can be decreased to accommodate more users. Of course, this requires more equipment, but this approach is cost effective on a per user basis, especially with the advances in technology.

2. New Uses—The Trend to Digital

a. Digital voice security. The general procedure for radio communications has been to send all voice transmissions via analog signals using one of the standard modulation types, AM, FM, or single sideband. While countries do have laws directed against the use of intercepted information by unintended recipients for personal gain, it is nearly impossible to keep people who have suitable receivers from listening.

Much of the information now transmitted is considered to be sensitive and so the trend is to encrypt the signal transmissions. The most secure encryption method is to convert all the transmitted information to digital form and then scramble the digits.

Digital voice security is used in the second generation (i.e., digital) cellular phones.

b. Data transmissions. In the fixed and mobile services, a rapid increase in the transmission of nonvoice or data messages is taking place. The new users of this service include computers transferring files among themselves. The trend is to do this at higher and higher information rates, which require wider bandwidth communication channels. This trend is changing the pattern of system usage as well as creating a need for more spectrum for these services.

c. Shared frequency bands. In many locations, microwave systems have been using the same frequencies that are used by satellite systems. This is possible because of improvements in directional antennas and careful allocation procedures.

Another method for sharing frequency bands is to use a coded modulation system known as spread spectrum. This system causes the transmitter power to be uniformly spread over a very wide band of frequencies, 1 to 50 MHz. The spread spectrum receiver, which is tuned to the desired signal code, receives other signals as noise. Thus it is possible to use spread spectrum

in the same frequency band with other types of radio systems.

d. Digitized time sharing. Telephone network designers have found that it is more efficient to digitize the communications and combine many users' communications into a single high-speed serial digital signal stream. In this type of system, a time segment corresponds to a frequency channel in a conventional radio system. In telephone systems, it is less expensive to switch, combine, separate, and direct digital signals. Although the bandwidth of this digital technique is greater than the combined bandwidths of the individual communications, this method, now used in some existing radio systems, may be more widely used in the future.

The use of this technology is now common in the fixed service bands where systems can be allocated the required wide channels. It is also used in digital Cellular and PCS phones that incorporate the Group Special Mobile (GSM) or the Time Division Multiple Access (TDMA) technology.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SYSTEMS, CIVILIAN • MICROWAVE COMMUNICATIONS • NETWORKS FOR DATA COMMUNICATION • RADIO PROPAGATION • SATELLITE COMMUNICATIONS • SATELLITE RF COMMUNICATIONS AND ONBOARD PROCESSING • SIGNAL PROCESSING, DIGITAL • TELECOMMUNICATIONS • VOICEBAND DATA COMMUNICATIONS • WIRELESS COMMUNICATIONS

BIBLIOGRAPHY

Broadcasting (1999). "Broadcasting and Cable Yearbook," R. R. Bowker, New Providence, NJ.

Couch, L. (1997). "Digital and Analog Communication Systems," 5th ed., Prentice-Hall, Upper Saddle River, NJ, ISBN 0-13-522583-3.

Federal Communication Commission (FCC) (1998), "Code of Federal Regulations," Vol. 47, Telecommunications, Superintendent of Documents, Pittsburgh, PA.

ITU General Secretariat (1998). "Radio Regulations," Edition 1998, Vols. 1–4, ISBN 92-61-07611.



Satellite Communications

Kenneth M. Peterson
Motorola, Inc.

I. Introduction
II. Satellite Systems
III. Communication Systems
IV. Link Budgets
V. Transmission Impairments

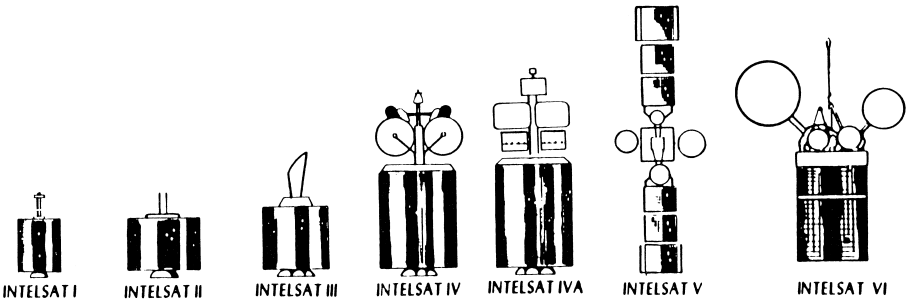
GLOSSARY

Apogee The point in the orbit of a satellite when it is most distant from the earth (or another planet).
Code division multiple access (CDMA) Process that shares time and frequency among many users by assigning to each an orthogonal code pattern that uniquely identifies each users signal.
Equivalent isotropic radiated power (EIRP) Measure of the power radiated by an antenna in the direction of a receiver expressed as the equivalent power that would have to be radiated uniformly in all directions.
Frequency division multiple access (FDMA) Process that shares a spectrum of frequencies among many users by assigning to each a subset of frequencies in which to transmit signals.
Perigee The point in the orbit of a satellite when it is nearest the earth (or other planet).
Time division multiple access (TDMA) Process that shares the time domain of a single carrier among many users by assigning to each time intervals in which to transmit signal bursts.

SATELLITE COMMUNICATIONS systems have been undergoing radical change in recent years, transitioning from a technology dominated by governments and geostationary satellites to one that includes Low Earth Orbit (LEO) and Medium Earth Orbit (MEO) satellite systems operated by publicly held corporations. These new systems utilize multiple antenna beams forming cells on the earth's surface similar to that used by terrestrial cellular telephony systems and are capable of carrying various types of traffic ranging from voice to Internet traffic.

I. INTRODUCTION

For many years, satellite communications systems were the exclusive domain of governments or international consortiums such as INTELSAT established by worldwide treaty. INTELSAT, an international communications satellite consortium managed by COMSAT between 1962 and 1978, and since then by INTELSAT's Director General and his staff, has been a world leader in commercial satellite system development. These satellites are illustrated in Fig. 1.



	INTELSAT I	INTELSAT II	INTELSAT III	INTELSAT IV	INTELSAT IVA	INTELSAT V	INTELSAT VI
TRANSPONDERS	2	1	2	12	20	27	48
BW/TRANSPONDER (MHz)	25	130	225	36	32, 36	40, 80, 240	40, 80, 160
COVERAGE	N. HEMISPHERE	GLOBAL	GLOBAL	GLOBAL & SPOT BEAMS	GLOBAL & SPOT BEAMS	GLOBAL, REGIONAL, & SPOT BEAMS	GLOBAL, REGIONAL, & SPOT BEAMS
e.i.r.p. (dBW)	11.5	15.5	23	22.5 (GLOBAL) 33.7 (SPOT BEAM)	22 (GLOBAL) 29 (SPOT BEAM)	22 - 29 (4 GHz) 44 (11 GHz)	23 - 31 (4 GHz) 41 - 44.4 (11 GHz)
NO. OF TEL.CIRCUITS	240 (NO MULT. ACCESS)	240	1200	4000(AVG.)	6000 (AVG.)	12000	30000
LIFETIME (YR)	1.5	3	5	7	7	7 +	10
SATELLITE COST/CIRCUIT YR ($K)	30	10	2	1	1	0.9	0.5

FIGURE 1 The INTELSAT satellites from 1964 to 1986.

Its first two commercial satellites, INTELSAT I and INTELSAT 11, used omnidirectional antennas, which while inefficient did serve to initiate multidestinational international communications with considerably greater ease than with either submarine cable or high-frequency radio. Each of these satellites accommodated as many as 240 telephone circuits by using frequency division multiplexed, multiple-access (FDMA) C-band carriers with FDM-FM modulation to achieve a fully interconnected network between earth stations located in North America and Europe. As satellites evolved, the power transmitted toward the earth (EIRP) and bandwidth were increased. INTELSAT III, with increased EIRP due to demployment of the first earth coverage antenna and employing two 250-MHz transponders, achieved increased capacity but still was power limited. INTELSAT IV channelized the transponders to 40-MHz bandwidth to increase power density, and in addition to earth coverage beams, used two 4-GHz, $3\frac{1}{2}^{\circ}$ spot beams to enhance down-link EIRP by about 8 dB but was capacity limited by intermodulation products and the total spectrum available. INTELSAT IVA was the first satellite to use multiple beams to achieve spectrum reuse by antenna pattern control.

In recent years, private corporations have been encroaching on this government-dominated arena by proposing a host of satellite communications systems intended to provide service to certain target markets. A total of more than 60 such systems have been proposed to meet the growing demand for international communications services. Both regional and global systems are contemplated, providing services ranging from mobile telephony to high-speed Internet access anywhere in the world. Examples of such systems include IRIDIUM, GLOBALSTAR, ICO, ELLIPSO, ORBCOMM, SPACEWAY, ASTROLINK, SKYBRIDGE, and TELEDESIC. Figure 2 shows the IRIDIUM communication system.

Each satellite in the IRIDIUM system, which uses a LEO constellation, forms 48 cells on the earth's surface by simultaneously generating 48 beams using a phased array antenna operating at L-band. The satellites communicate with each other via crosslinks utilizing a carrier frequency of approximately 23 GHz. The IRIDIUM system is designed to carry both voice and data traffic in addition to providing paging services.

The different types of proposed systems are each in various stages of development. Each system, either planned or existing, has a unique configuration optimized to support a business plan based on the services offered and the markets targeted. These satellite systems may be either regional systems or global systems. Regional systems, which tend to utilize Geostationary Orbits (GEO), are aimed at serving a particular geographic market. Some of the global systems intend to provide global telephony service, filling in where terrestrial wireless systems leave off or providing seamless connectivity between different terrestrial systems. Others intend to provide global data connectivity, either for low-cost short-message applications such as equipment monitoring or for high-speed Internet access.

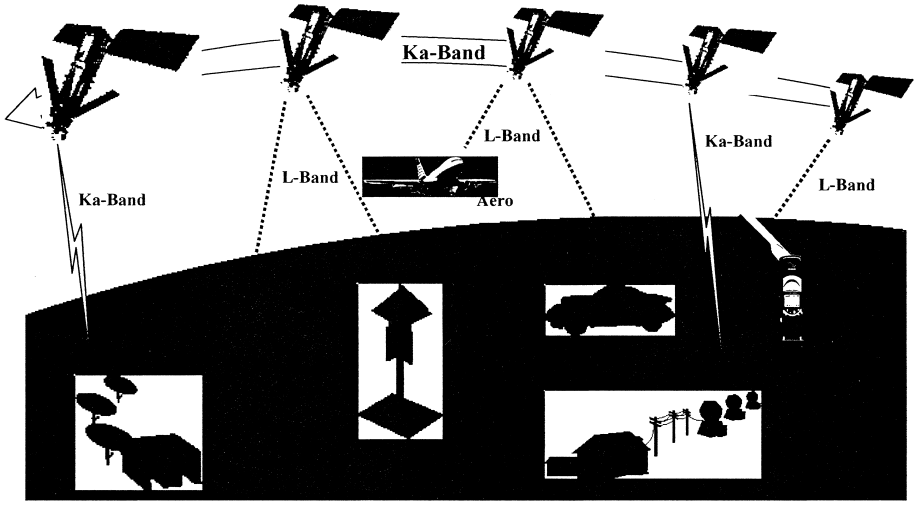


FIGURE 2 The IRIDIUM satellite communication system.

Another important potential market for these systems is that part of the world that is unserved or underserved by existing terrestrial infrastructure.

Most of the proposed satellite communications systems are intended to serve markets that are largely untested, and it is likely that economics, user acceptance rates, technical difficulties, and other factors will cause adjustments in the business plans of many of these systems.

II. SATELLITE SYSTEMS

Communications satellites are radio-relay stations in space. They serve much the same purpose as the microwave towers one sees along the highway. The satellites receive radio signals transmitted from the ground, amplify them, and retransmit them back to the ground. Since the satellites are at high altitude, they can "see" across much of the earth. This gives them their principal communications advantage: the ability to span large distances. Ground links, such as microwave relays, are inherently limited in their ability to cover large distances by the terrain.

Satellite communications systems consist of several different segments including a space segment, a ground control segment, and ground infrastructure equipment. The space segment contains the satellites, which act as nodes in space routing the communication signal from an earth terminal to either a final or intermediate destination point on the earth's surface. The ground control segment is responsible for monitoring the health and status of each satellite as well as for keeping the satellites in their appropriate locations in space. The infrastructure equipment controls the networking aspect of the entire communication system, keeping track of the duration of a communication session for billing purposes as well as assigning communication channels to the various users.

A. Orbits

The selection of an orbit for a satellite communications system is closely tied to the market it is intended to serve and the type of communications offered. There are four different orbit types that may be utilized: GEO, LEO, MEO, and Highly Elliptical Orbit (HEO). Each of these orbits has various characteristics, which make it more or less desirable for the designated market. One factor that influences the choice of orbit is space debris, which is present in varying, amounts at different altitudes as shown in Fig. 3.

Space debris can, depending on the orbit, result in the destruction of an entire plane of satellites. Thus an altitude must be chosen so that debris is not a significant threat. The selection of an orbit must also include consideration of other aspects of the space environment such as operation in the energetic charged particle environment. The energetic charged particle environment is often termed the space radiation environment due to the effects, which result from the impingement of these particles upon the spacecraft surfaces and the traverse of them through spacecraft materials and components.

Three sources of energetic charged particles compose this environment: particles trapped in the earth's geomagnetic field, particles traversing interstellar and interplanetary space, and those emanating from the sun. The Van Allen radiation belts consist of energetic protons and electrons, which become trapped in the geomagnetic field of the earth, as depicted in Fig. 4. The inner belt region consists of both electrons and protons; the outer belt region contains energetic electrons.

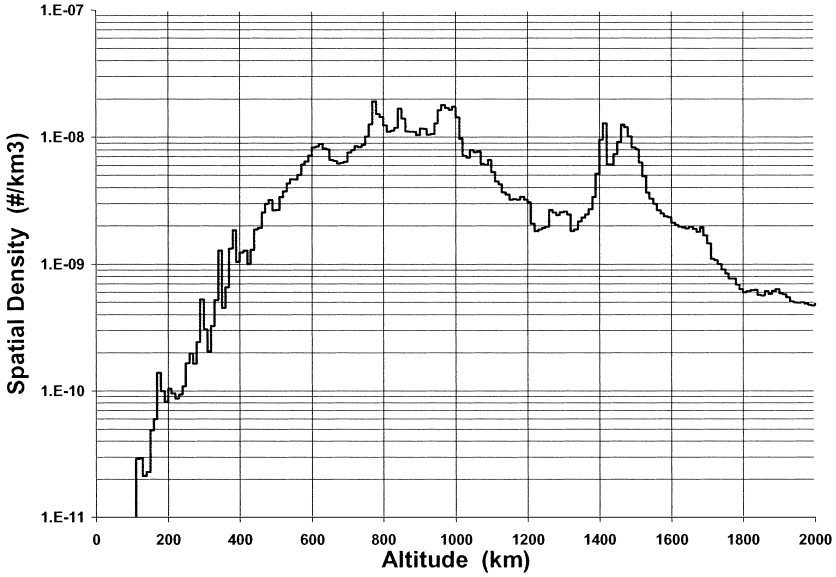


FIGURE 3 Space debris at different altitudes as of January 1998.

1. GEO

GEO satellites orbit the earth at a fixed distance of 35,786 km. The speed of the satellite at this distance matches the earth's rotation, thereby keeping the satellite stationary over a particular point on the earth. The majority of communication satellites to date are in geostationary orbit, and this is expected to remain the case for some time. Geostationary satellites support voice, data, and video services usually providing service to a particular region. A majority of U.S. long distance companies utilize geostationary satellites to provide back-up capacity for telephony service. Examples of GEO satellite systems include INTELSAT, Inmarsat, and PanAMSat.

The geostationary orbit typically provides coverage from 20 degrees north latitude to 20 degrees south latitude. Geostationary satellites orbit the earth above the equator with a single satellite capable of providing coverage over approximately one-third of the earth's surface. GEO systems are known for their reliability and operational predictability, and have the advantage of being able to provide revenue to the service operator upon the launching of a single satellite.

Satellites are placed in geostationary orbit by first being launched into a transfer orbit as shown in Fig. 5.

A geostationary satellite carries with it a solid-fuel rocket engine to provide the necessary "kick" at the apogee of the satellite's transfer orbit to place it in a circular equatorial orbit with essentially zero inclination. The apogee engine and its propellant weigh about as much as the rest of the satellite components put together.

The geostationary orbit is disadvantaged for certain types of services such as voice because of the relatively large time delay (about 0.25 sec) experienced by a communication signal in traveling from the earth to a geostationary satellite and back. This delay is sufficiently long to be objectionable in long distance telephone calls that make use of a GEO satellite. Other services such as video conferencing also find long time delays to be objectionable. For this reason, GEO systems are best utilized where time delay is not an issue as is the case with direct TV service and Internet data services.

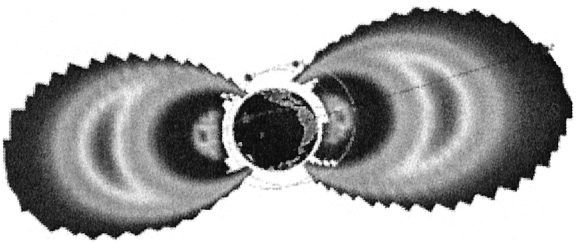


FIGURE 4 The Van Allen radiation belts.

2. LEO

LEO systems orbit the earth at altitudes ranging from about 700 to 3000 km, and unlike GEOs, move with respect to the earth. A typical LEO satellite orbits the earth in less than two hours, which means that a satellite is in view for only a few minutes. Because of this, users of LEO satellite services must be "handed off" from satellite to satellite as well as between "cells" within the footprint of a given satellite in order to maintain the continuity of

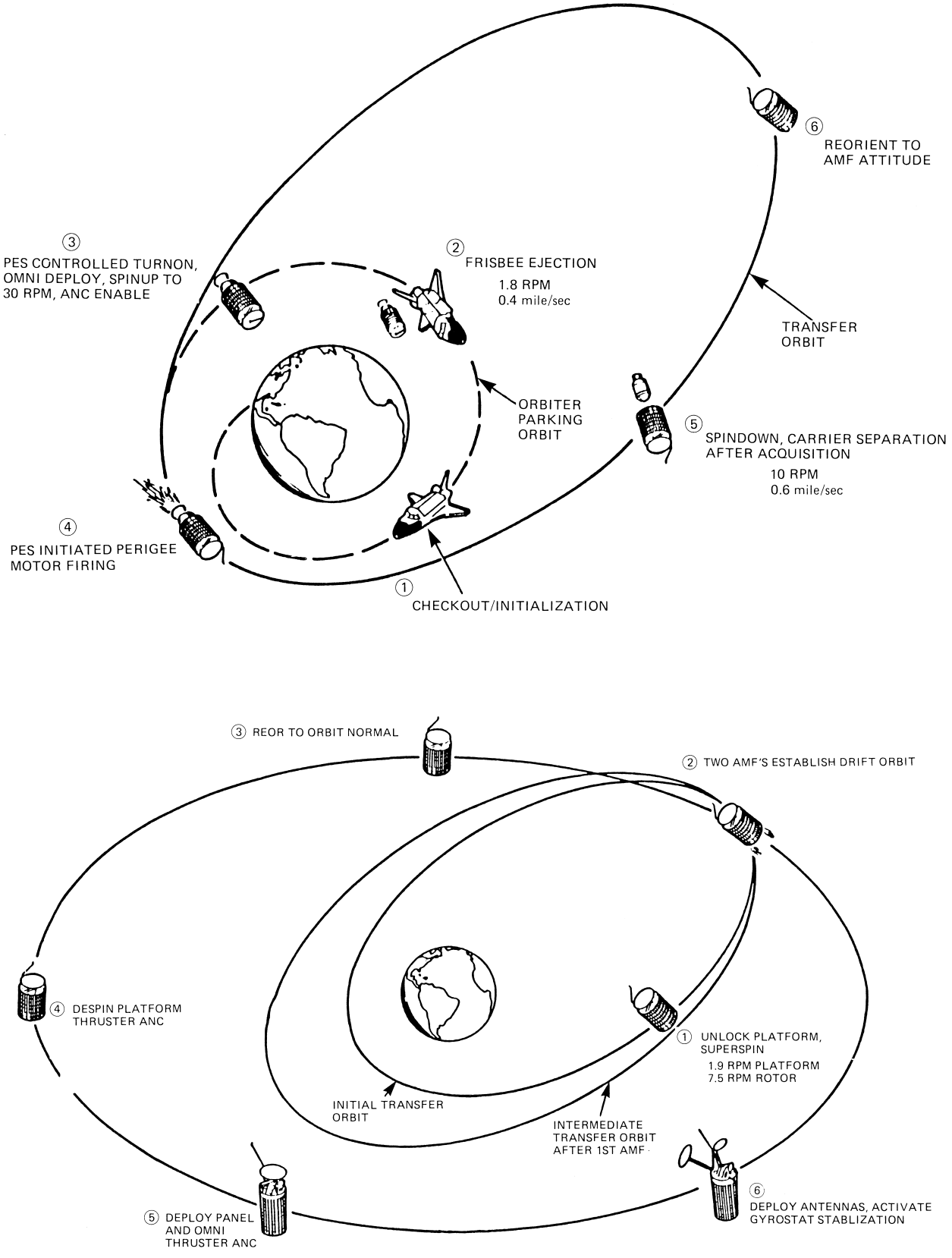


FIGURE 5 (a) The STS launch to transfer orbit and (b) transfer orbit to synchronous orbit.

communication. The handoff between cells of a particular satellite is usually accomplished within the satellite itself; however, handoffs between satellites requires either the intervention of a ground terminal or communication between the satellites using crosslinks or some hybrid of these two approaches. Also, because of the relative velocity between a satellite and any particular point on the earth, Doppler is present on each of the communication links. In most cases, this Doppler must be corrected for in the earth terminal equipment.

The choice of altitude for a LEO constellation is usually governed by the cost of placing satellites at the particular altitude selected. The higher the altitude, the fewer the number of satellites required to obtain full earth coverage, but it is usually more costly to launch satellites to higher altitudes. Figure 6 shows the number of satellites required to obtain full earth coverage as a function of altitude.

Other factors that influence the choice of altitude are the radiation environment as a function of altitude and the amount of space debris at any particular altitude. The chosen altitude is generally a compromise between several often conflicting factors, not the least of which is the intended market or mission.

A good example of a LEO system is the IRIDIUM system. This system was intended to service voice and low-speed data users, and orbits the earth at an altitude of 780 kilometers. At this altitude, debris is not a significant

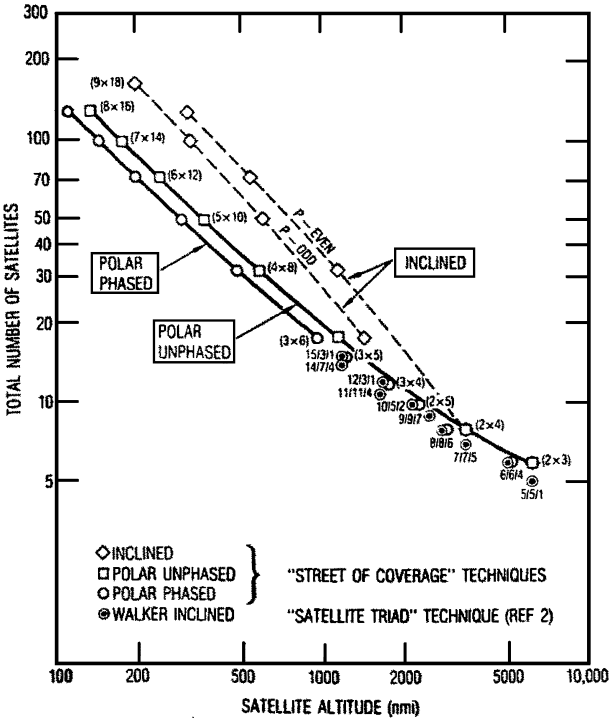


FIGURE 6 Number of satellites for full earth coverage as a function of altitude.

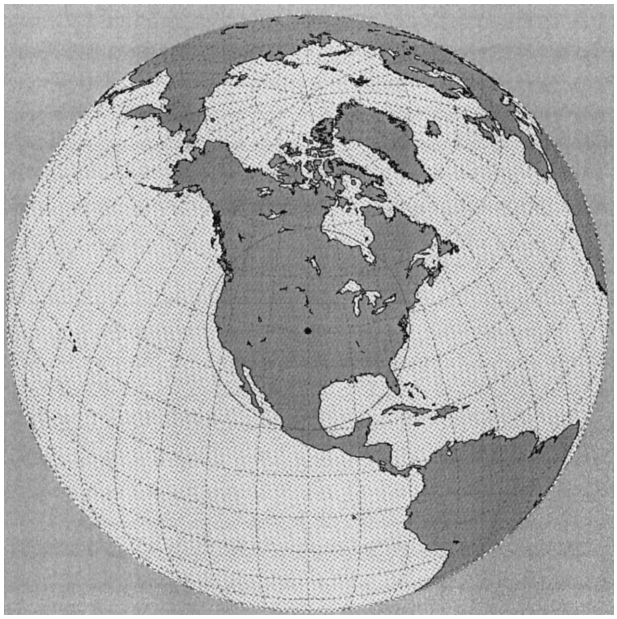


FIGURE 7 Footprint of an IRIDIUM satellite.

factor, and the radiation environment is relatively benign. Figure 7 shows the footprint of an IRIDIUM satellite at the designed minimum elevation angle of 8.2°.

LEO systems are disadvantaged in that essentially all of the constellation's satellites must be in place before revenue can begin to flow for commercial systems. This is in sharp contrast to GEO systems, where revenue may be had with a single satellite in place. Also, the ground control equipment needed to maintain a LEO constellation is more extensive than that required for a GEO system since ground sites must be located at several points over the earth to ensure adequate control.

The advantage of LEO systems is that the satellites' proximity to the ground enables them to communicate with minimal time delay. Thus for services that are delay sensitive such as voice communication, these constellations are ideal. Further, the shorter distance to the earth means that the satellite to earth communication links suffer less path loss, and hence a reliable link can be established with less power and or reduced antenna size. This usually results in LEO satellites being smaller and of lower mass and hence less expensive than their GEO counterparts.

3. MEO

MEO systems operate at an altitude of approximately 10,000 km above the earth, which is an orbit that provides a compromise between the characteristics of the GEO and LEO orbits. Compared to LEOs, the higher altitude requires fewer satellites to provide full earth coverage and provides a much longer view time from a point on the earth

to the covering satellite. The relative velocity between a MEO satellite and the earth is also less, so the resulting Doppler present on the space to earth communication links is less than for LEO constellations operating at the same frequency band. Compared to GEO systems, MEOs can operate effectively with smaller, mobile equipment and with less time delay on the communication links. One of the disadvantages of MEO systems is that the radiation environment is quite intense for these orbits, and hence more radiation hardening is necessary in the electronic subsystem which in turn translates into higher cost satellites.

Most MEO constellations consist of 10 or more satellites distributed over two or three orbital planes. They can be used to provide an array of services ranging from voice to Internet data. An example of a MEO system is the ICO Global Communication system.

4. HEO

HEO systems operate differently than GEOs, LEOs, or MEOs. In a HEO constellation, the satellites orbit the earth traversing an elliptical path rather than a circular one. In such an orbit, the satellite velocity with respect to the earth is constantly changing, reaching a maximum at perigee and a minimum at apogee. The footprint of the satellite on the earth is constantly changing as well, reaching a maximum diameter at apogee. Although these HEO systems can be designed to give full earth coverage, they are usually designed to cover those regions of the earth where the satellite dwell time is a maximum.

Current proposed systems that employ HEO orbits are actually hybrid systems in that they also include a GEO satellite or a set of MEO satellites orbiting the equator. A good example of a HEO system is the ELLIPSO system, which is intended to provide both voice and data services. The proposed ELLIPSO constellation utilizes six MEO satellites orbiting the equator in addition to two HEO planes of satellites.

B. Satellites

A satellite is made up of several subsystems including the mechanical structure, the thermal subsystem, a power subsystem, payload, propulsion system, and an attitude control system. Satellites are designed to meet the specific requirements associated with the mission or consumer market they are intended to service. Their design is influenced by the characteristics of available launch vehicles, the characteristics of the spacecraft payload, and the chosen orbit.

Satellites designed for use in GEO systems, for example, are considerably different than those intended for LEO systems (Fig. 8), as GEO satellites require a great deal more onboard propellant to reach their final altitude. Also, the payload of a GEO satellite will typically employ a much larger main mission antenna structure than a LEO satellite designed for a similar mission.

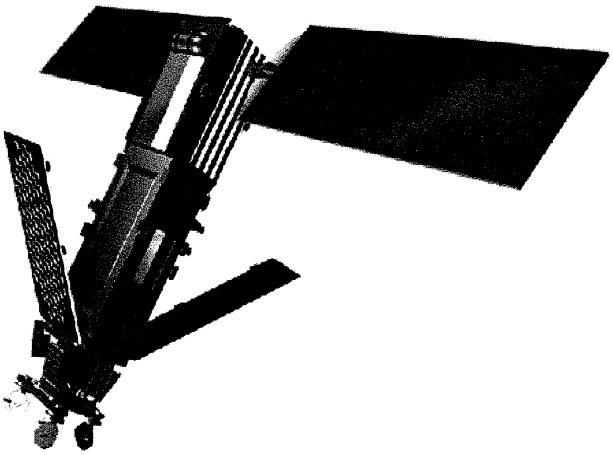


FIGURE 8 IRIDIUM satellite.

To maintain its attitude properly, some satellites are spin stabilized, so that they maintain their orientation in space by means of a rotation of the body of the satellite along an axis parallel to the earth's axis (Fig. 9), and others are platform stabilized and maintain their orientation by small jets and momentum wheels (Fig. 10).

All are powered by arrays of silicon or gallium arsenide solar cells mounted on the body. The satellite body serves as a platform for the communications payload. In the case of a spin-stabilized satellite, this platform is "despun" in order to maintain its earth orientation.

The stabilization system depends on infrared sensors to determine the direction of the earth and on small thrusters to align the satellite axis properly and keep the antenna beams pointed earthward. For orbital station keeping, it is necessary to minimize the satellite's tendency to drift along its orbit because of the gravitational perturbations due to the oblateness of the earth.

The satellite is carefully tracked by ground control stations, the propulsion vector needed to compensate for drift is computed, and the appropriate thrusters are fired by radio command. The thrusters, fueled by hydrazine, generate the reaction force necessary to adjust the satellite's position. With them the stabilization system can control the attitude of the platform, move the satellite within the orbital plane, and counteract the gravitation influence of the sun and moon, which tends to perturb the orientation of the satellite. The correction for orientation consumes a major fraction of the hydrazine fuel and is therefore closely related to the effective life expectancy of the satellite.

It is also necessary to monitor the condition of the various systems and devices in the satellite, and to take

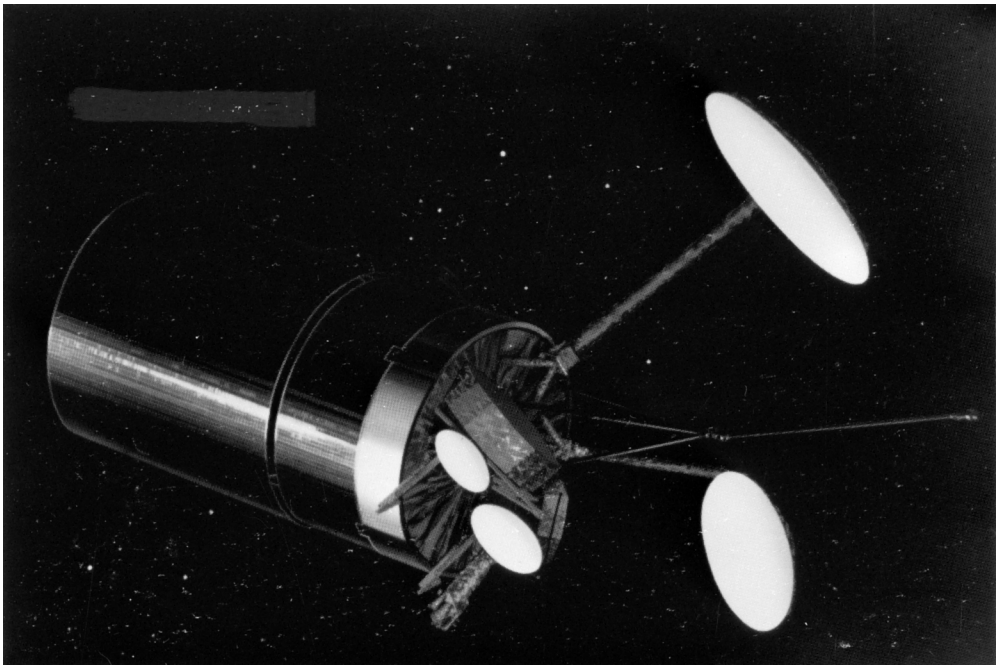


FIGURE 9 INTELSAT VI spin-stabilized communications satellite.

corrective action in the event of a malfunction. This is accomplished through telemetry in which a variety of sensors detect temperatures, voltages, and other conditions that are transmitted to ground control stations through a radio link. The information is subsequently sent to the control center, where it is processed for evaluation and displayed. If any out-of-tolerance condition is detected in the satellite, corrective action can be taken by radio command from the same control stations.

The effectiveness of a communications satellite is measured not only by its lifetime but also by its capacity, which in turn depends on radiated power and bandwidth. As the technologies improve, satellites will continue to improve in all three areas: lifetime, power, and bandwidth.

C. Earth Stations

To support a satellite constellation and its mission, the ground system must provide a command and control function for each of the spacecraft and furnish the necessary infrastructure for the communications network. These two functions generally require separate types of ground equipment, as is the case for example in the IRIDIUM system. To support LEO constellations, the command and

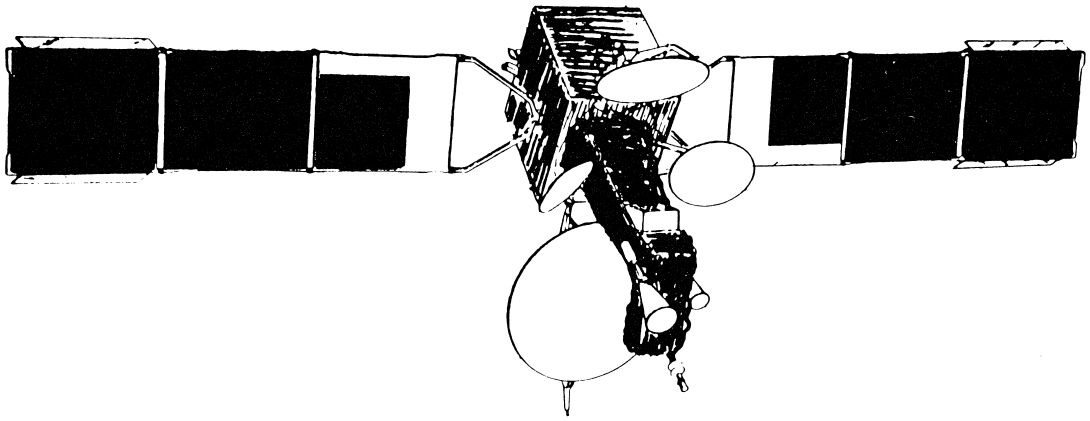


FIGURE 10 INTELSAT V stabilized platform satellite.

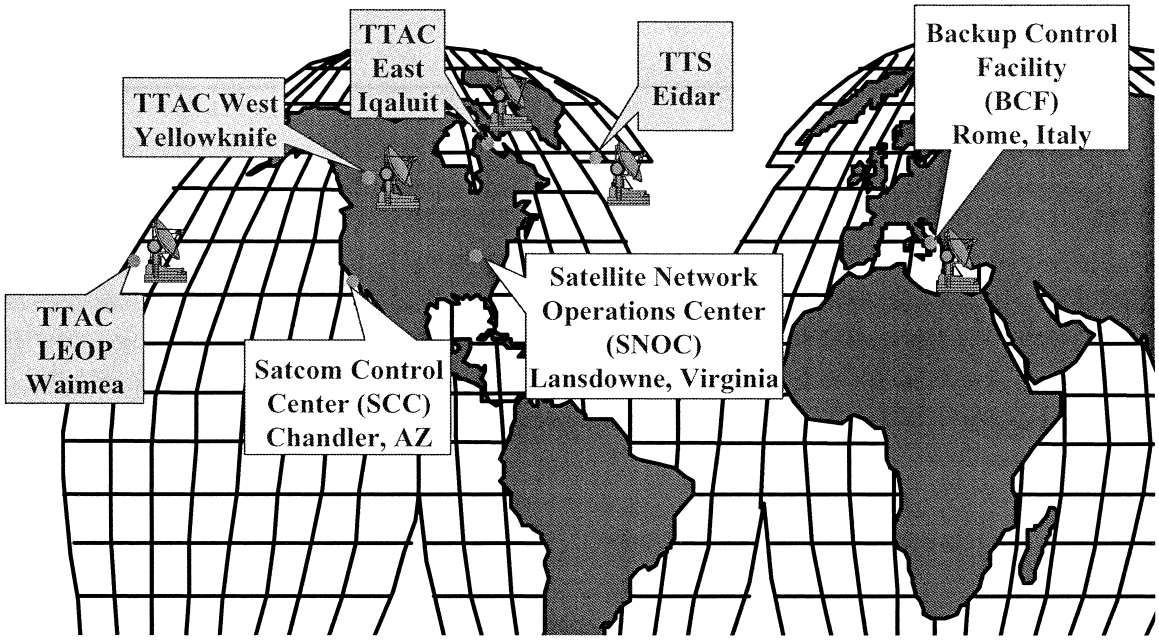


FIGURE 11 The IRIDIUM command and control system.

control function requires several transmit/receive terminals generally located in the northern latitudes as shown in Fig. 11 for the IRIDIUM system.

The satellite operations center monitors the health and status of each of the satellites in a constellation, tracks the satellites to determine orbital position, and determines spacecraft attitude from sensor information. The satellite operations center controls the spacecraft as well as status of the payload by transmitting command data to the spacecraft. The ground system uses spacecraft housekeeping telemetry and mission data to carry out these functions. As an example, the control center may use data from a spaceborne radar altimeter to refine knowledge of the spacecraft's orbit.

The antennas used in the earth station equipment are usually large diameter parabolic reflectors as seen in the INTELSAT ground station shown in Fig. 12.

The large antennas are required to transmit powerful radio signals to a satellite and receive weak signals from it. The power needed for transmitting the signal to the

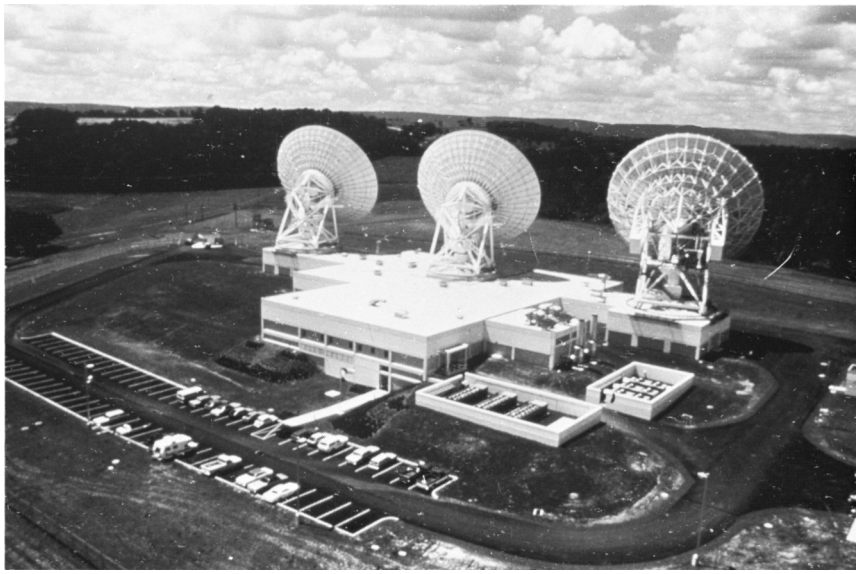


FIGURE 12 Earth station for international satellite communications at Roaring Creek, Pennsylvania.

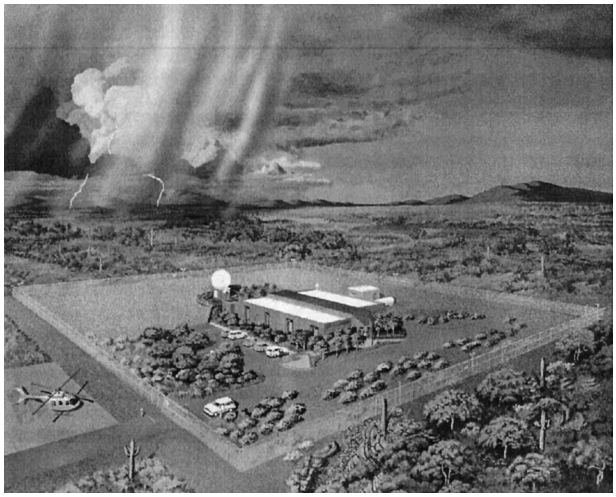


FIGURE 13 IRIDIUM Gateway.

satellite is developed by an amplifier that can produce from 1 Watt to several kilowatts depending on system requirements before it is fed to a parabolic antenna ranging from 1.8 to 30 m in diameter, depending on the application. The function of the antenna is to confine the radiated energy beam aimed at the satellite. The value of power radiated by the antenna is the product of antenna gain and high-power amplifier (HPA) power, and is referred to as the equivalent isotropically radiated power.

Ground station equipment, sometimes referred to as "gateways," control the communications network and interface the satellite communication system to existing terrestrial networks. Figure 13 shows an artist's rendition of a typical gateway used in the IRIDIUM system.

The gateway equipment controls the network traffic, provides billing services for the network operator, records certain details of the communication session including session start and termination times, and keeps records pertinent to the subscribers associated with a particular gateway operator. If the satellite system is providing global services as opposed to regional, then gateways are usually located throughout the world, as shown in Fig. 14 for the IRIDIUM system.

The receivers, transmitters, and antennas associated with the gateway equipment are very similar to that in the ground control equipment often operating in the same frequency band. With non-GEO systems, the earth station antennas must track the overhead satellite from horizon to horizon employing a second antenna to pick up the oncoming satellite as the first satellite disappears from view. This provides a "make before brake" connection so that communication between the earth station and the satellite constellation is uninterrupted.

Since both the ground control and gateway equipment functions are so critical, it is imperative that the communication links to the constellation operate with very high "availability." Since rain can be an impairment to these links (this depends on the operating frequency), often multiple antennas are used to improve the link availability. These antennas are so spaced as to extend beyond the diameter of typical rain cells. During rain, the link between the satellite and earth station can then be switched to that antenna not affected by the rain. Additionally, the links will have added margin, i.e., added power will be transmitted to overcome the increased losses encountered when a link operates in rain. Both of these factors

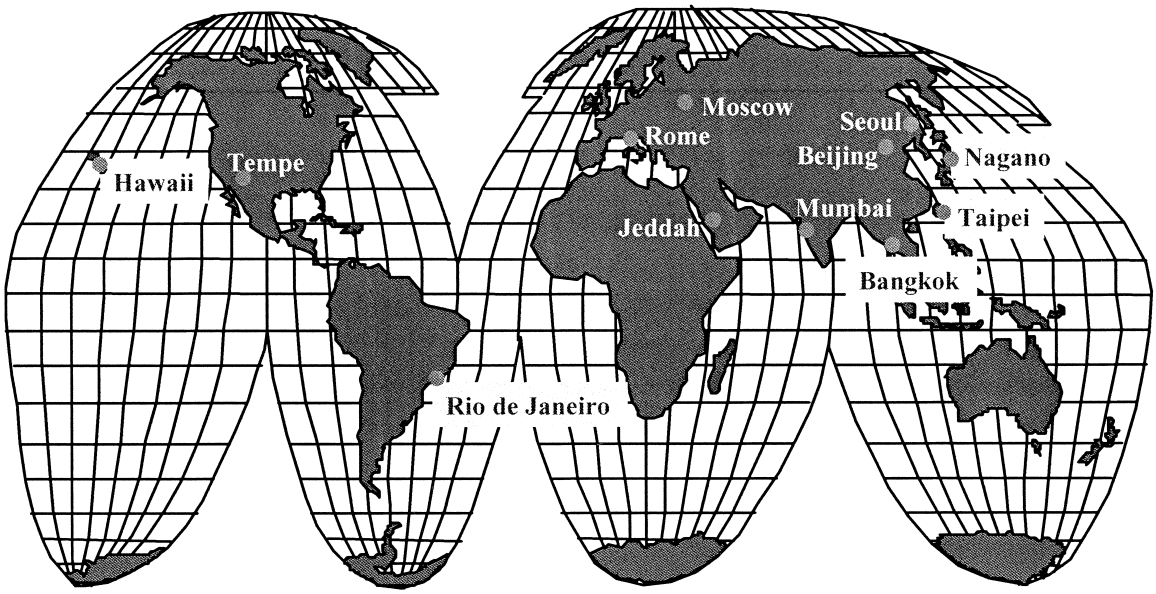


FIGURE 14 Initial IRIDIUM Gateway locations.

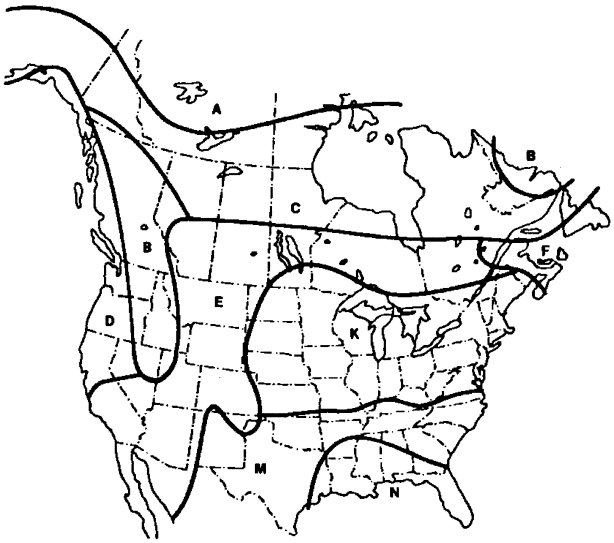


FIGURE 15 ITU rain regions for the Continental United States and Canada.

improve link availability substantially. Typical link availability ranges from 0.99 to 0.99999, depending upon the system. Figure 15 shows the International Telecommunications Union (ITU) rain regions over the United States and Canada (letters farther along in the alphabet designate the higher rain rates, i.e., region K has a higher rain rate in mm/hr than region D, etc.).

III. COMMUNICATION SYSTEMS

The communication architecture of a satellite system includes the satellite payload communication subsystem, satellite crosslink communication subsystem, and the communication subsystem linking the satellite with the ground control as well as the ground infrastructure segments. The satellite payload communication subsystem is primarily responsible for providing communication to the intended market be it a handheld cellphone or subscriber fixed site equipment. To fulfill this mission, the satellite payload utilizes multiple beam antennas, be they fixed parabolic reflectors or phased arrays, to form cells on the earth's surface. In a GEO system these cells are fixed on the earth's surface; however, with LEO or MEO systems the cells are constantly moving over the earth. Figure 16 shows a typical cell structure for a single IRIDIUM satellite in which 48 simultaneous cells are formed.

For LEO or MEO systems, depending upon the market application, some of these cells may have to be turned off as a satellite traverses from equator to pole and then turned back on again once the satellite is over the pole. This is necessary because as a satellite traverses from equator to pole, the footprint begins to overlap that of neighboring satellites in adjacent planes. Figure 17 shows the resulting cell pattern for the IRIDIUM system as generated by an onboard cell shutdown algorithm.

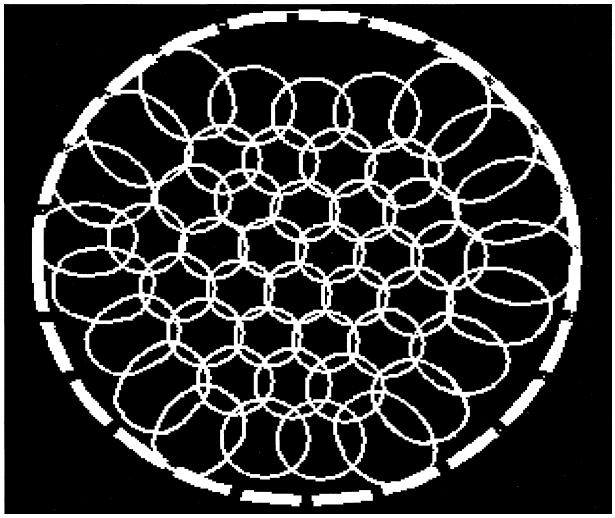


FIGURE 16 IRIDIUM satellite cell structure.

The beam coverage pattern for a GEO system may be very different from that for a LEO, or MEO, as illustrated by the beam structure for the INTELSAT V regional GEO satellite shown in Fig. 18.

A. Regulatory Considerations

Satellite communications frequency allocations, and regulations, fall under the influence of a combination of three types of entities: an administration, a group of

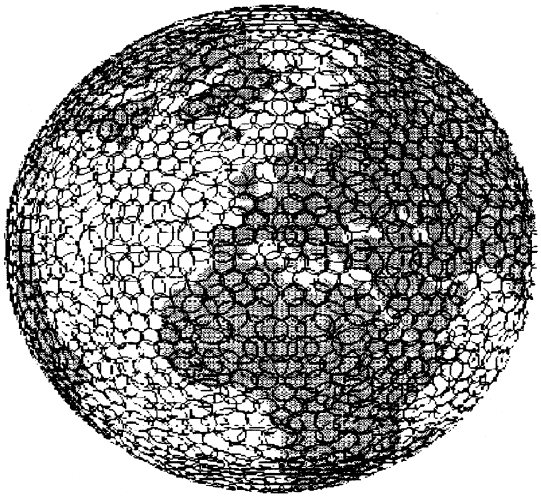


FIGURE 17 IRIDIUM constellation cell pattern.

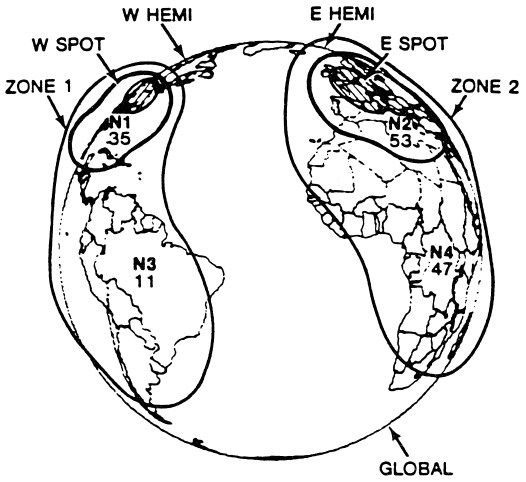


FIGURE 18 INTELSAT V beam coverage.

administrations, or an international body. Each administration (or nation) has a sovereign right to specify the conditions of operation that a satellite system must follow in order to provide service within the border of that administration. Since the region of service of a satellite system typically encompasses many administrations, often they form groups to work together in order to develop rules/regulations for these systems to ensure a common regulatory regime. One example of this type of group is the European Conference of Postal and Telecommunications Administrations (CEPT), which is developing harmonized allocations throughout Europe. The final group, the ITU, is an international body that considers all regions of the world and gathers all administrations to develop a common set of allocations.

The ITU is affiliated with the United Nations (UN) organization. It implements the treaty that contains the international allocation of frequencies. These allocations are incorporated in the Radio Regulations. Changes to these international allocations are developed and modified at World Radio Conferences (WRC) that meet every two to three years, which are treaty conferences. As a result of a WRC, changes are proposed to the Radio Regulations, and once ratified by each administration, these changes are treated like any other treaty. The ITU table of allocations, which are part of these regulations, specifies the bands of frequencies for different services in three distinct regions of the earth as shown in Fig. 19.

Table I shows a sample of the frequency allocations for satellite systems operating in the 24.75–29.9 GHz band.

B. Transmission

Each of the various communications subsystems associated with a satellite system utilizes some type of architecture to accomplish the intended function. For example, the payload communication subsystem might use a transponding architecture, as shown in Fig. 20.

Alternatively, it may employ a processing architecture in which the communication signals are demodulated onboard the spacecraft and remodulated again on a different carrier frequency for transmission (Fig. 21).

The particular payload architecture used will depend upon the intended market or mission as well as the satellite orbit.

In a similar manner, the satellite crosslink communication subsystem and the ground control and infrastructure subsystems utilize an architecture that consists of antennas, transmitters, receivers, and any auxiliary equipment required to facilitate communication. Most of these subsystems utilize solid state electronics, including the transmitter power amplifier, to achieve the desired system reliability and service availability. Modern satellite crosslink subsystems may even use laser communications to achieve

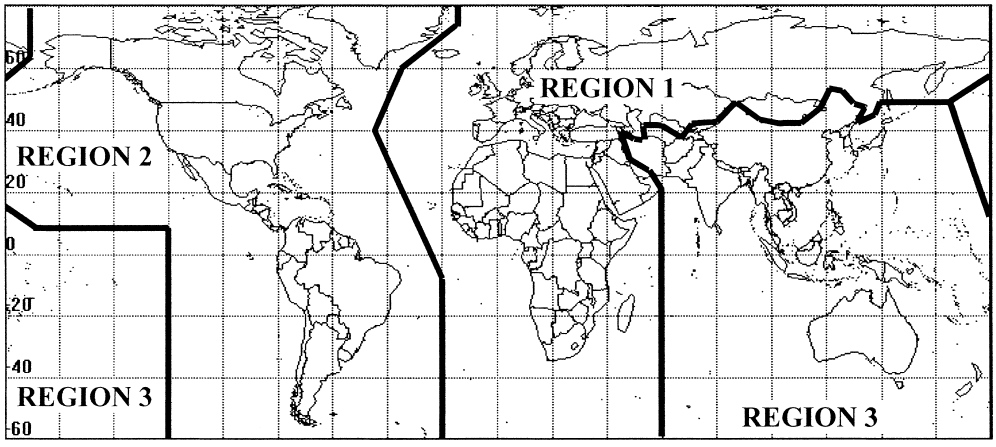


FIGURE 19 ITU regulatory regions.

TABLE I 24.75–29.9 GHz Band

Allocation to services		
Region 1	**Region 2**	**Region 3**
24.75–25.25 Fixed	24.75–25.25 Fixed satellite (earth to space) S5.535	24.75–25.25 Fixed Fixed satellite (earth to space) S5.535 Mobile S5.534
25.25–25.5 Fixed	Intersatellite S5.536 Mobile Standard frequency and time signal satellite (earth to space)	
25.5–27 Earth exploration satellite (Space to earth) S5.536A S5.536B	Fixed Intersatellite S5.536 Mobile Standard frequency and time signal satellite (earth to space)	
27–27.5 Fixed Intersatellite S5.536 Mobile	27–27.5 Fixed Fixed satellite (earth to space) Intersatellite S5.536 S5.537 Mobile	
27.5–28.5 Fixed	Fixed satellite (earth to space) S5.484A S5.539 Mobile S5.538 S5.540	
28.5–29.1 Fixed	Fixed satellite (earth to space) S5.484A S5.523A S5.539 Mobile Earth exploration satellite (earth to space) S5.541 S5.540	
29.1–29.5 Fixed	Fixed satellite (earth-to-space) S5.523C S5.523E S5.535A S5.539 S5.541A Mobile Earth exploration satellite (earth to space) S5.541 S5.540	
29.5–29.9 Fixed satellite (earth to space) S5.484A S5.539 Earth exploration satellite (earth to space) S5.541 Mobile satellite (earth-to-space)	29.5–29.9 Fixed–satellite (earth to space) S5.484A S5.539 Mobile satellite (earth to space) Earth exploration satellite (earth to space) S5.541 S5.525 S5.526 S5.527 S5.529 S5.540	29.5–29.9 Fixed satellite (earth to space) S5.484A S5.539 Earth exploration satellite (earth to space) S5.541 Mobile satellite (earth to space)
S5.540 S5.542	S5.542	S5.540 S5.542

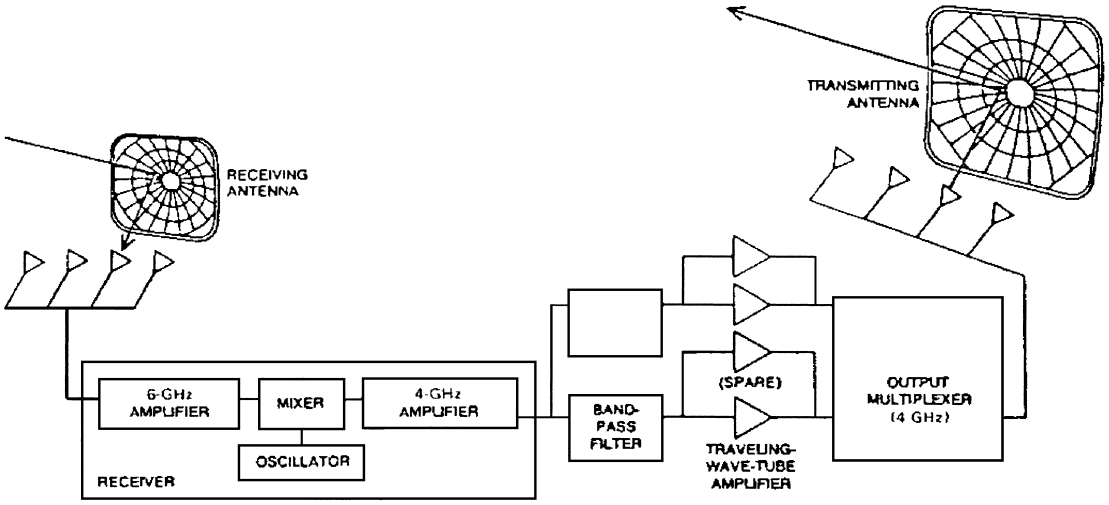


FIGURE 20 Satellite payload transponding architecture.

the required information transmission rates. Depending on the market to be served, these rates can be as high as several gigabits per second.

1. Modulation

Numerous modulation techniques are used in accomplishing signal transmission over satellite links. While older satellite communications systems have used amplitude and frequency modulation techniques, more recent systems utilize digital modulation. Digital modulation techniques extensively used for supporting digital satellite transmission are BPSK (binary phase shift keyed) and QPSK (quadrature phase shift keyed). BPSK can be expressed as

$$B(t) = \cos(2\pi f_c t + i\pi),$$

where i is the binary state sequence of 0's and 1's representing the digital signal. QPSK can be expressed as

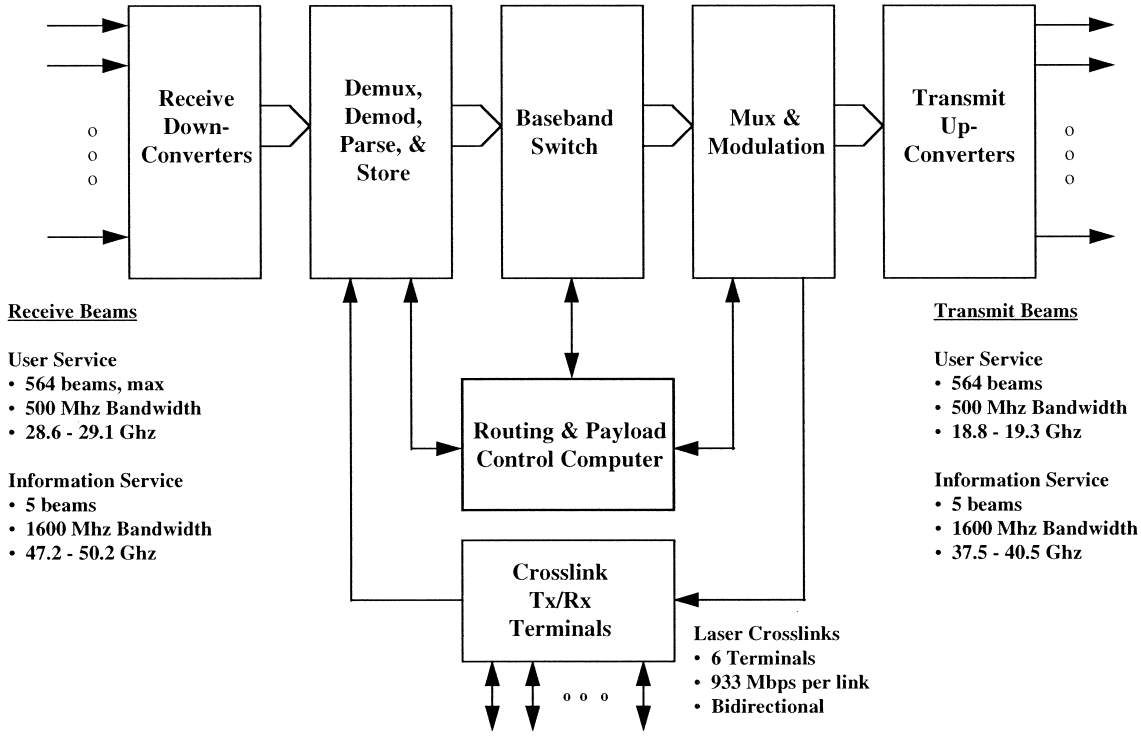


FIGURE 21 Satellite payload processing architecture.

$$Q(t) = \cos(2\pi f_c t + j\pi/2),$$

where j is determined by the states of pairs of bits in the bit sequence according to the table

Pair	j
00	0
01	1
11	2
10	3

The bits are transmitted in terms of symbols of duration T. In the case of BPSK, each symbol period of the modulated carrier identifies one bit in terms of carrier phases of 0 and π, while in the case of QPSK each symbol period identifies two bits in terms of carrier phase angles of $0, \pi/2, \pi$, and $3\pi/2$. The signal occurring in each symbol period is a sinusoid burst of frequency f_c and duration T having a power spectrum distribution given by the relation

$$S(f) = \{\sin[\pi(f - f_c)T]/[\pi(f - f_c)T]\}^2.$$

This power spectrum theoretically extends from minus infinity to plus infinity; however, the Nyquist sampling theorem shows that the entire information content of the digital stream is contained in a bandwidth of $W = 1/T$. In practice, the bandwidth may be made slightly greater (about 20%) to allow for imperfections in realizing the ideal filter. BPSK and QPSK are used extensively to carry satellite digital communications from rates of a few bits per second to several hundred megabits per second. In certain applications where the objective is to carry a greater density of communications in the frequency spectrum, 8 PSK modulation has been considered; however, because of its sensitivity to phase distortion, coding must be introduced to maintain low bit error rate. A number of other modulation techniques have also been introduced, each with the intention to improve the quality of digital communications via satellite principally by improved resistance to the influence of nonlinearities encountered in the high power amplifiers used in the earth stations and on-board the satellites. Some of these are offset QPSK, frequency shift keyed (FSK), minimum frequency shift keyed (MSK), serial MSK (SMSK), and quadrature amplitude modulation (QAM). Also, both BPSK and QPSK can be modulated in terms of the phase difference between symbols, a technique that avoids the necessity to recover the carrier to perform demodulation. In this case they are called DBPSK and DQPSK, respectively. Many other modulation techniques are possible; however, at present the phase shift keyed techniques are the most highly utilized.

C. Access Technologies

It is common practice to divide the radio spectrum assigned to satellite service into subbands, which are referred to as channels, and these are used to communicate to and from a spacecraft. In a satellite equipped with a transponding form of payload, each transponder receives weak uplink signals arriving from ground stations, amplifies them, and retransmits them back to earth. Transponder channels at C-band frequencies typically are spaced nominally 40 or 80 MHz apart, and have a useful bandwidth of 36 or 72 MHz, respectively. At K band, transponder channels with frequency spacings ranging from 40 to 250 MHz are in use.

The frequency plan of the transponders used on the INTELSAT V communications satellite is shown in Fig. 22. A satellite will carry enough transponders to occupy the full radio frequency spectrum space assigned to it. The 500-MHz spectrum assigned to fixed satellite communications at the C-band can accommodate twelve 40-MHz transponders or six 80-MHz transponders. By frequency reuse of the same spectrum, accomplished by using different polarizations and spacial separations in the form of beams, the available spectrum can be multiplied. For example, INTELSAT V reuses the 500-MHz C-band four times.

Each transponder receives signals sent to it on its up-link channel frequency (in the 6–GHz C band or in the 12-GHz Ku-band) from many earth stations, translates these to the appropriate down-link channel frequency (in the 4-GHz C-band or in the 2-GHz Ku-band), and after amplification, retransmits them back to the earth.

The signal sent from any one of the earth stations on the up-link can be received by all of the stations in the down-link. This is commonly referred to as point-to-multipoint or multidestinational operation, and is an important property unique to satellite systems that cannot be easily matched by terrestrial telecommunications services and is responsible for the great utility of satellites in telephone and data telecommunications applications. Also, a multiplicity of earth stations can share each transponder channel by transmitting signals on different frequencies or at different times or with different codes. These different ways of multiple accessing are referred to as frequency-division multiple access (FDMA), time-division multiple access (TDMA), or code-division multiple access (CDMA), respectively.

1. FDMA

Each earth station accessing the satellite with FDMA is assigned its own exclusive carrier frequency and bandwidth within a particular transponder channel on both the up-link and down-link along with the carriers of several other

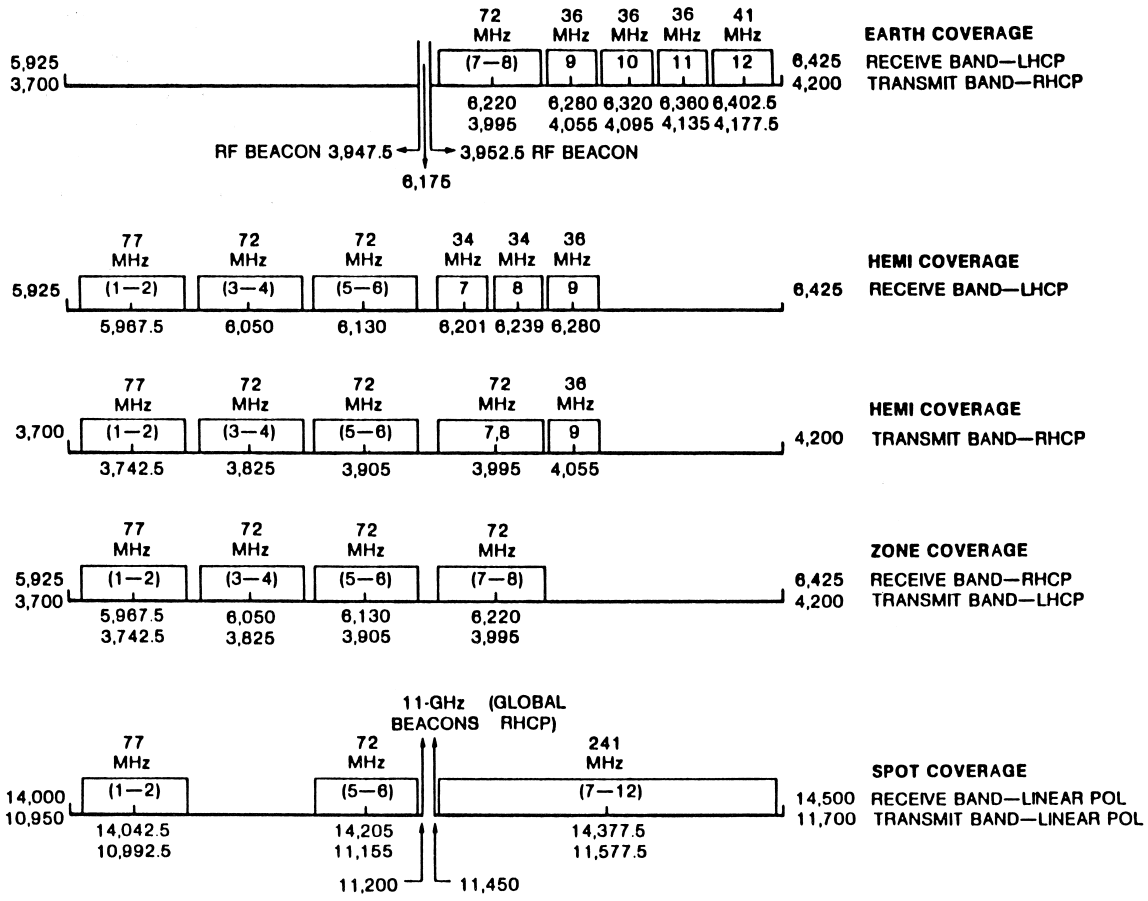


FIGURE 22 INTELSAT V transponder center frequency bandwidths and polarizations.

stations. Any station wishing to receive another station's transmitted signal will tune its receiver to that station's down-link frequency. Stations accessing a transponder with FDMA will be assigned different bandwidths according to their traffic capacity. The carrier frequencies must be adjusted to ensure that the bands occupied by each carrier do not overlap the bands of the other carriers and cause excessive adjacent carrier interference. Because of this requirement, any readjustment of the capacity assigned to one of carriers requires corresponding adjustments to the others and can entail a considerable coordination effort. Typically, the power amplifiers used in the transponder are nonlinear, and because FDMA operates with several carriers sharing a transponder channel, distortion products due to intermodulation create noiselike interference, which limits the fraction of power that can be used. Transponders operating with FDMA are typically backed off to 40% or less of their full power output to achieve near linear operation and reduce the interference. Even so, FDMA has been the dominant method for the delivery of satellite services during the first two decades of satellite communications, and is expected to continue to be so in the future. FDMA is used for carrying analog trunk telephone on both domestic and international satellites, thin route single-channel-per-carrier (SCPC) telephony and data, demand-assigned multiple access (DAMA) SCPC, broadcast television distribution for major television networks, and cable television distribution; it is currently being introduced for international digital trunk telephone distribution. Typically, the capacity achieved for FDMA operation in a 40-MHz transponder channel carrying several carriers in the INTELSAT system is 500-analog telephone channels. This increases to 1000-analog telephone channels if a single carrier occupies the transponder.

2. TDMA

Each earth station accessing the satellite with TDMA occupies its own exclusive fraction of time (time slot) in a time frame (TDMA frame) shared with several other stations that occupy other time slots. A typical TDMA frame is shown in Fig. 23.

The TDMA frame is periodic with a period that may range from as short as 125 μsec to as long as 30 msec. During its assigned time fraction, a station transmits a short

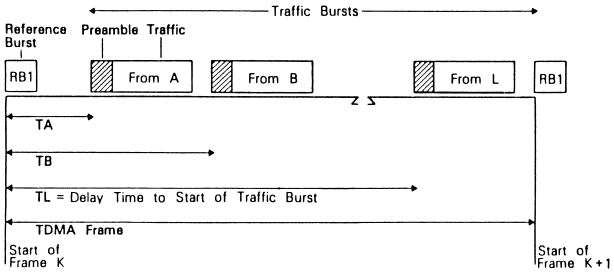


FIGURE 23 TDMA frame organization.

burst at an assigned carrier frequency that contains its traffic in time compressed form. On the transmit of a TDMA terminal, the signal presented to the input during a period equal to the TDMA frame is stored in a memory and is transmitted at an assigned time as a short burst in the next TDMA frame. Because this memory compresses the time interval containing the signal, it is called the compression memory. Usually, the compression memory consists of two memories operating in a ping-pong fashion, in which one stores the current input signal while the other transmits the previous frame's and in the next frame the roles are reversed and so on. Any station wishing to receive a station's transmitted signal will adjust its receiver to open a time window at the expected time of arrival of the transmitting station's traffic burst in the TDMA frame. On the receiver side of the terminal, the received burst is stored in a memory operating at the high transmission rate and is read from this memory at the normal signal rate during the next frame. Because this memory expands the time interval of the short burst back to its normal duration, it is called the expansion memory. It is usually implemented in the same ping-pong fashion as the compression memory. All stations transmit their traffic bursts on the same carrier frequency using a high signaling rate that completely fills the transponder channel. Typically, the signaling rate is 60 Mbit/sec for 40-MHz transponders and 120 Mbit/sec for 80-MHz transponders. The high bit rate carrier is shared by several stations accessing the TDMA frame in their assigned time fractions. An input signal arriving at a rate of 64 Kbit/sec and served by a terminal operating at a transmission rate of 60 Mbit/sec will be transmitted in a burst having a duration that is $T/938$, where T is the duration of the TDMA frame period. (Note that the transmission time is compressed by the ratio of the transmission rate to the input signal rate.) A few additional bits are added to the transmitted burst as a preamble to provide information needed for carrier frequency and bit timing recovery, acquisition, synchronization, burst position control, and network control. TDMA traffic bursts may be either preassigned, in which case they maintain an assigned location and duration in the frame for long periods of time, or demand assigned, in which case the position and duration of a burst may be varied each time a new call arrives or an old call departs the system. Because each receiving station must have a timing reference with which to gauge the time of arrival of other stations' traffic burst in the TDMA frame, the TDMA frame contains a reference burst that is transmitted from a reference station. The reference station may be one of the traffic stations that is nominated to be the reference or it may be a special station intended to perform only the reference station function. For redundancy, some systems have two reference bursts.

The traffic carrying capacity of a TDMA station is adjusted simply by changing the duration and location of its traffic burst. A process called a coordinated burst time plan change has been devised for TDMA systems, which permits the reallocation of capacity among stations to be carried out synchronously by changing the time locations and durations of the traffic bursts of all stations in the same TDMA frame without interruption of ongoing traffic. The key to performing such synchronous burst time plan changes is to establish a fixed round trip time delay equal to an integer number of TDMA frames between the time a reference burst leaves the satellite and the time a response from any earth station synchronized to that burst returns to the satellite. This time delay, called the coordination time, is the same for all stations and is selected to be greater than the round trip delay between the satellite and the farthest station on the earth. To accomplish burst time plan synchronization, each station inserts a time delay, D_N, between the time a reference burst marking the start of a received TDMA control multiframe is received and the start of the station's transmit control multiframe. This value is selected to be equal to the difference between the actual round-trip delay for the station and the coordination time. The relationship for determining D_N is

$$D_N = MT_F - 2d_N/c,$$

where T_F is the duration of the TDMA frame, M is the number of TDMA frames sufficient to cause the product MTF to exceed the round-trip propagation time between the satellite and the farthest station, d_N is the actual distance between satellite and station N, and c is the speed of light. The derivation of the above expression is illustrated in Fig. 24.

D_N is a function of the station's distance from the satellite, and is continuously adjusted by feedback or open loop control to assure that the coordination round-trip delay is always maintained. Because the maximum distance between the satellite and an earth station is 26,000 miles, the coordination time must be at least 280 msec.

When TDMA systems use only one carrier per transponder in multiple-access satellite service, they avoid

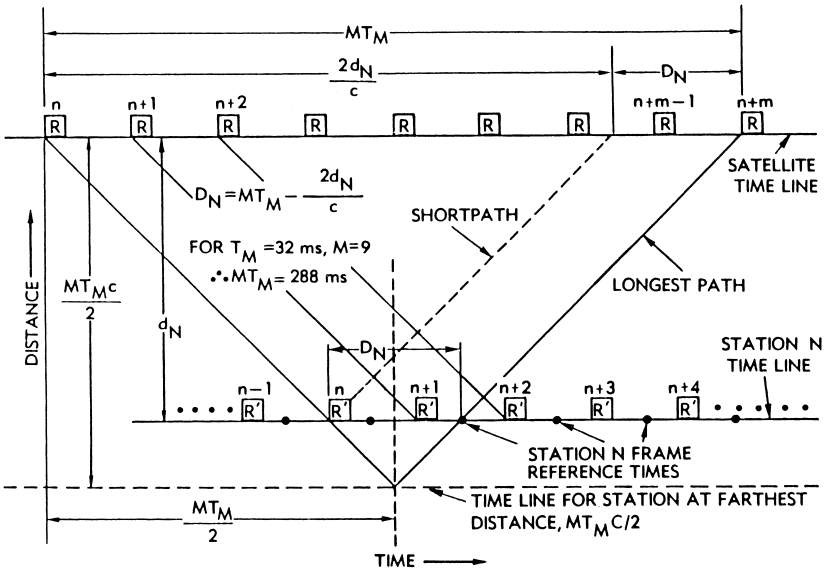


FIGURE 24 Time–space graph for determination of synchronization delay D_N, for traffic station N.

the intermodulation distortion problem encountered with FDMA. They can operate very near the saturated power output of the traveling wave tube (TWT) used in the transponder provided that care is taken in selecting the filters used through the system. In the INTELSAT TDMA system, the TWT can be operated 1 dB below its saturated power output. In the SBS TDMA system, the TDMA carrier channel filters are significantly wider than the TDMA carrier s bandwidth. This permits the TWT to be operated several decibels in excess of saturation without serious impairment, a condition that is very desirable in combating up-link fades at its K_u-band operating frequency.

When TDMA is used in low rate applications and several TDMA carriers are carried in the same transponder, the space segment power advantage of single-carrier-per-transponder TDMA is lost because of multiple carrier intermodulation distortion. This results in loss of space segment capacity; however, such low bit rate operation greatly reduces the size of the earth station antenna or power amplifier because the power radiated from the earth terminal's antenna is typically reduced by approximately 8–10 dB and the channel noise is reduced due to the reduced channel bandwidth.

TDMA systems operating in the single-carrier-per-transponder mode achieve very high traffic carrying density. The capacity of a single 80-MHz transponder in the INTELSAT system using a 120-Mbit/sec TDNIA carrier is conservatively 16,000 full-time 64-kbit/sec digital PCM telephone channels, which with digital speech interpolation is increased to 3200 telephone channels. The 64-kbit/sec PCM technique for carrying telephone signals is now being replaced by a 32-kbit/sec adaptive differential pulse coding modulation (ADPCM) technique referred to as low rate encoded (LRE), which when used jointly with digital speech interpolation will double the telephone channel capacity, thus yielding 6400 telephone channels per transponder. Systems operating in the low bit rate FDMA/TDMA manner achieve approximately half of the single-carrier-per-transponder capacity, but the stations using the system are significantly smaller, lower in power, and lower in first cost and operations and maintenance costs.

3. CDMA

In a CDMA system, the individual bits of a message are encoded in terms of unique digital codes. Many stations, each having its own unique code, can simultaneously transmit their signals and each can be discriminated from all of the others. The discrimination is accomplished by means of a cross correlator, which uses as a reference the code corresponding to that of the desired transmission. Each transmitted symbol of binary information is expressed in the form of a coded sequence of shorter duration symbols called chips. The number of chips per symbol G is an important CDMA parameter referred to as the processing gain or bandwidth spreading factor because the bandwidth of the transmitted signal is expanded by the factor G. In a typical application for 9.6-kbit/sec transmission, a code having 7000 chips per symbol may be used. Thus, the 9.6-kbit/sec signal, which when transmitted in conventional quarternary phase shift keyed (QPSK) form requires a bandwidth of approximately 4.8 kHz,

when coded for CDMA will require a QPSK bandwidth of 33.6 MHz. Because of this increased bandwidth imparted by CDMA, it is frequently referred to as a spread spectrum technique.

The codes used for CDMA are specially selected to have good cross-correlation properties so that interference from simultaneously present but unwanted transmissions of other signals can be significantly reduced. A family of psuedorandom (PR) sequence codes called GOLD codes are considered to have suitable properties. A desirable code is one that, when multiplied by itself, symbol by symbol, and summed over all of its symbols, yields an output N, where N is the number of symbols in the code, but when shifted yields very near $N/2$. Each receiver uses a unique PR code; hence, a station wishing to correspond with a particular receiver simply transmits the message using the destination receiver's code. In order to receive a desired CDMA signal, a receiver stores the PR code of the desired signal into its cross correlator and correlates it with the total received signal. When the chips of the received desired signal align with the stored code, the cross-correlator puts out a positive or negative impulse depending whether a binary 0 or 1 was transmitted. The cross-correlation process is able to select the desired signal code even though it is buried in the composite of many other coded transmissions and thermal noise. Implementation of the cross-correlator is typically accomplished by means of a surface acoustical wave (SAW) correlator.

The number of simultaneous transmissions that can be supported using CDMA is limited by the self-generated random signal background caused by the sum of the individual signal transmissions plus the system's thermal noise background. Each unwanted signal arriving with carrier power C will contribute a noiselike interference of C/G. Thus, in a system having M simultaneous arriving signals of which only one is wanted, this noiselike interference will be MC/G. To this must be added the thermal noise contribution N. If the objective received bit energy to noise ratio Is E_b/N_0, then for QPSK modulation, the number of simultaneous received signals must not exceed the limit given by the expression

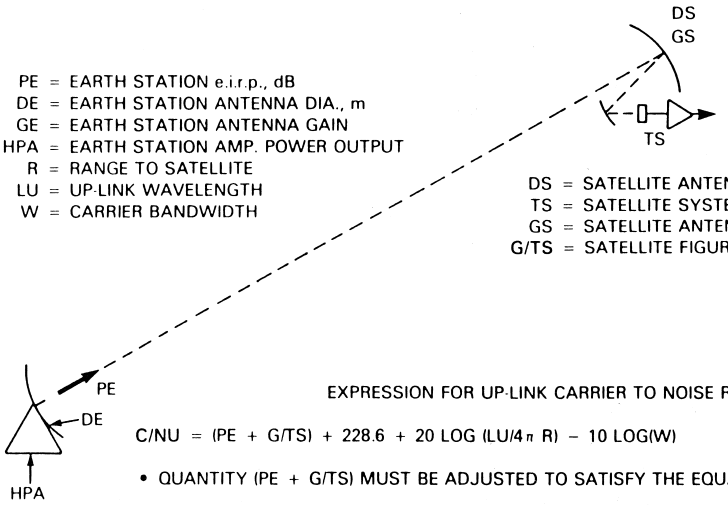
$$M \leq G\left[(2E_b/N_0)^{-1} + (C/N)^{-1}\right].$$

For a bit error rate of 10^{-6} requiring an E_b/N_0, of 12 dB for a practical demodulator with 1.6 dB of implementation margin and assuming a carrier to thermal noise ratio C/N of 20 dB, the limit is $G/131.7$. For the example, considered previously using $G = 7000$, the limit is 291 simultaneous signals. Since each signal can support 9.6 Kbit/sec, the system can carry a total bit rate of 2.79 Mbit/sec. Introduction of channel coding can improve the total data capacity of the system, as can be illustrated by the following example. Assume that each message is rate 1/2 encoded, which yields a 5-dB reduction in the E_b/N_0 needed to achieve the same bit error rate (BER) objective. To accommodate the doubling of the rate per channel, the value of G must be halved to 3500. The net result is to increase the number of simultaneous signals to 384 and the total channel capacity to 3.69 Mbit/sec. An attending advantage of CDMA resides in the action of the receive side correlator processing, which causes the undesirable impact of narrow band interfering signals to be greatly reduced.

CDMA can be used to access a satellite system in two different ways. One is to identify each station in terms of its transmitted code, in which case a receiving station must use a separate correlator for each station it wishes to receive. The other is to identify each station in terms of its reception code, in which case each transmit station must transmit on different codes for each receiver. This requires only one receive side correlator. One application of CDMA used in satellite systems is to transmit low data rate messages from a centrally located station to many remote, small diameter, receive only stations. In this case CDMA is used because of its capability to selectively address the remote stations by use of different codes and to significantly reduce the interference from other satellite systems by virtue of the spectrum spreading property. Another application of CDMA currently being used is in the GLOBALSTAR system, which is intended for voice and data transmission to mobile as well as fixed location users.

IV. LINK BUDGETS

The various parameters involved in calculating satellite link budget are illustrated in Fig. 25. Let the power transmitted in the direction of the satellite from the earth station be designated as P_E. This is determined from the product of antenna gain in the direction of the satellite and the power delivered to the antenna feed flange from the earth station's high-power amplifier (HPA). This results in a power flux density of $P_E/4\pi R_U^2$ at the satellite, where R_U is the distance from the earth station to the satellite. Assuming that the satellite has a gain G_S in the direction of the transmitting earth station and the transmission is at a wavelength L_U, the satellite receives an up-link carrier power level Cu, equal to $P_E G_S (L_U/4\pi R_U)^2$. The thermal noise on the up-link is given by the product of Boltzmann's constant k, the satellite system noise temperature T_S, and the carrier bandwidth W, which is $kT_S W$. Hence, the up-link carrier to thermal noise ratio when expressed in decibel form is



PE = EARTH STATION e.i.r.p., dB
DE = EARTH STATION ANTENNA DIA., m
GE = EARTH STATION ANTENNA GAIN
HPA = EARTH STATION AMP. POWER OUTPUT
R = RANGE TO SATELLITE
LU = UP-LINK WAVELENGTH
W = CARRIER BANDWIDTH

DS = SATELLITE ANTENNA DIA., m
TS = SATELLITE SYSTEM NOISE TEMP
GS = SATELLITE ANTENNA GAIN
G/TS = SATELLITE FIGURE OF MERIT

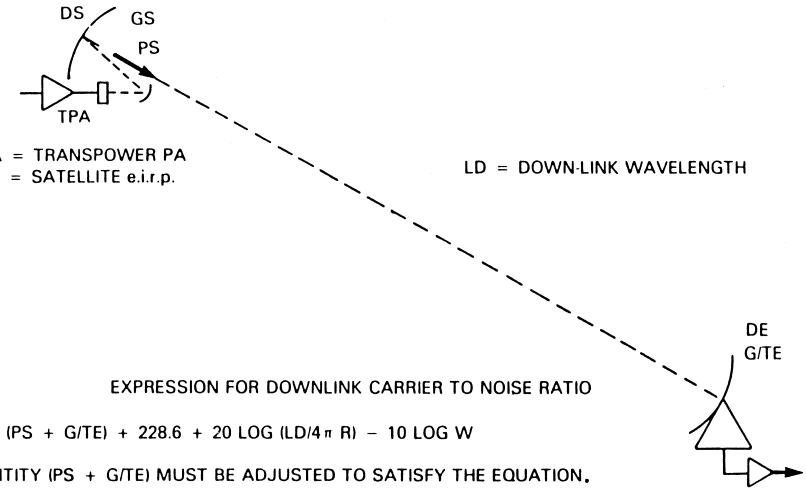
EXPRESSION FOR UP-LINK CARRIER TO NOISE RATIO

$$C/NU = (PE + G/TS) + 228.6 + 20 \log (LU/4\pi R) - 10 \log(W)$$

- QUANTITY (PE + G/TS) MUST BE ADJUSTED TO SATISFY THE EQUATION.
- MOST COMMERCIAL SYSTEMS TODAY USE HIGH PE AND LOW G/TS. THIS LEADS TO EXPENSIVE EARTH STATIONS.
- INCREASED SIZE OF SATELLITE APERTURE INCREASES G/TS, REDUCING PE AND HENCE EARTH STATION SIZE.
- INTERFERENCE, C/IU, MUST ALSO BE FACTORED IN

$$\therefore C/NUT = \left[(C/NU^{-1} + C/IU^{-1})\right]^{1}$$

(a)



TPA = TRANSPOWER PA
PS = SATELLITE e.i.r.p.

LD = DOWN-LINK WAVELENGTH

EXPRESSION FOR DOWNLINK CARRIER TO NOISE RATIO

$$C/ND = (PS + G/TE) + 228.6 + 20 \log (LD/4\pi R) - 10 \log W$$

- QUANTITY (PS + G/TE) MUST BE ADJUSTED TO SATISFY THE EQUATION.
- MOST COMMERCIAL SYSTEMS TODAY USE RELATIVE LOW PS AND HIGH G/TE. THIS LEADS TO EXPENSIVE EARTH STATIONS.
- INCREASED SIZE OF SATELLITE APERTURE INCREASES GS AND PS, REDUCING G/TE AND HENCE SIZE OF EARTH STATION.
- INTERFERENCE, C/ID, MUST ALSO BE FACTORED IN

$$\therefore C/NDT = \left[(C/ND^{-1} + C/ID^{-1})\right]^{-1}$$

(b)

FIGURE 25 (a) Satellite up-link budget calculation and (b) satellite down-link budget calculation.

$$(C/N)_{\mathrm{U}} = P_{\mathrm{E}} + 20\,\mathrm{Log}(L_{\mathrm{U}}/4\pi R_{\mathrm{U}}) + G/T_{\mathrm{S}}$$

Transmit power | Path loss | Satellite G/T

$$+ 228.6 - 10\,\mathrm{Log}W$$

Boltzmann's constant | Noise bandwidth

This result must be modified to account for the influence of up-link interference from other earth stations pointed at other satellites and from other beams of the same satellite and intermodulation distortion. This is done by calculating the carrier to up-link interference ratio, $(C/IF)_{\mathrm{U}}$, from knowledge of the up-link interference environment and the carrier to up-link intermodulation ratio, $(C/IM)_{\mathrm{U}}$, from the knowledge of the intermodulation distortion generated in the earth station's HPA. These interferences are then summed on a power basis with the thermal noise using the following relation expressed in decibel form:

$$(C/N)_{\mathrm{UT}} = -10\log\left(10^{-0.1(C/N)_{\mathrm{U}}} + 10^{-0.1(C/IF)_{\mathrm{U}}} + 10^{-0.1(C/IT)_{\mathrm{U}}}\right).$$

The same expressions can be developed for the down-link in terms of the following variables: P_{D} for the power radiated in the direction of the earth station from the satellite, L_{D} for the downlink wavelength, R_{D} for the down-link distance between the satellite and the earth station, G/T_{E} for the earth station's antenna gain to system noise temperature ratio, C/N_{D} for the down-link carrier power to thermal noise ratio, $(C/IF)_{\mathrm{D}}$ for the down-link interference from other beams on the same satellite plus adjacent satellites, and $(C/IM)_{\mathrm{D}}$ for the intermodulation distortion generated in the satellite transponder. The expressions for the down-link case are the same as those given above for the up-link with obvious variable substitutions. From the above relations, the individual up- and down-link carrier to thermal noise plus interference and intermodulation distortion ratios are determined. These are then combined using noise power addition to determine the overall end-to-end carrier to noise plus interference and intermodulation ratio. The expression for calculating this overall result in terms of the up-link and down-link contributions is

$$(C/N)_{\mathrm{UDT}} = -10\log\left(10^{-0.1(C/N)_{\mathrm{UT}}} + 10^{-0.1(C/N)_{\mathrm{DT}}}\right).$$

The value of $(C/N)_{\mathrm{UDT}}$ thus calculated is compared to that needed to accomplish the desired performance objective and the difference is the margin. It is typical to require an overall margin of 2 dB for C-band operation and 8 dB for K_u-band operation. It is convenient to develop a tabular format for calculation of link budgets. One example is shown in Table II.

TABLE II Satellite Link Budget Calculation

Up-link	
Pu, up-link radiated power	85.5 dB W
Up-link spreading loss, 4ir/R2u	163.4 dB/m^2
Flux density at satellite for 2.5 dB IBO	−77.9 dB W/m^2
Up-link path loss for $L_U = 0.05$ m, 6 GHz	200.4 dB
Satellite Rcv G/T	−14.5 dB/K
Boltzmann's constant	228.6 dB/Hz K
Noise bandwidth, $W = 30$ MHz for QPSK	74.8 dBHz
$(C/N)_{\mathrm{U}}$	24.4 dB
$(C/IF)_{\mathrm{U}}$	27.0 dB
$(C/IM)_{\mathrm{U}}$	∞ dB
$(C/N)_{\mathrm{UT}}$	22.5 dB
Down-link	
PD, satellite down-link power at beam edge	21.8 dB W
Satellite down-link path loss, $L_{\mathrm{D}} = 0.075$ m	197.0 dB
Earth station GIT, 30-m INTELSAT std A	39.0 dB/K
Boltzmann's constant	228.6 dB/Hz K
Noise bandwidth, $W = 30$ GHz	74.8 dBHz
$(C/N)_{\mathrm{D}}$	17.6 dB
$(C/IF)_{\mathrm{D}}$	27.0 dB
$(C/IM)_{\mathrm{D}}$	∞ dB
$(D/N)_{\mathrm{DT}}$	17.1 dB
End-to-end link	
$(C/N)_{\mathrm{UD}}$	16.0 dB
Objective, link BER = 10−′	12.5 dB
Margin	3.5 dB

V. TRANSMISSION IMPAIRMENTS

A. Path-Induced Impairments

The most prevalent form of path-induced impairments that influence satellite communications at the C- and K_u-band frequencies are those due to rain. Rain causes three types of path impairment to appear: attenuation, increased noise temperature, and scattering. Regarding attenuation, the presence of water in the path between the earth station and the satellite causes energy absorption due to the excitation of molecular resonances. These are present to some extent at the C-band frequencies and to a much large extent at the K_u- and K_a-band frequencies. Data on such absorption is plotted in terms of the probability of exceeding a given level of path attenuation for an assumed weather model in

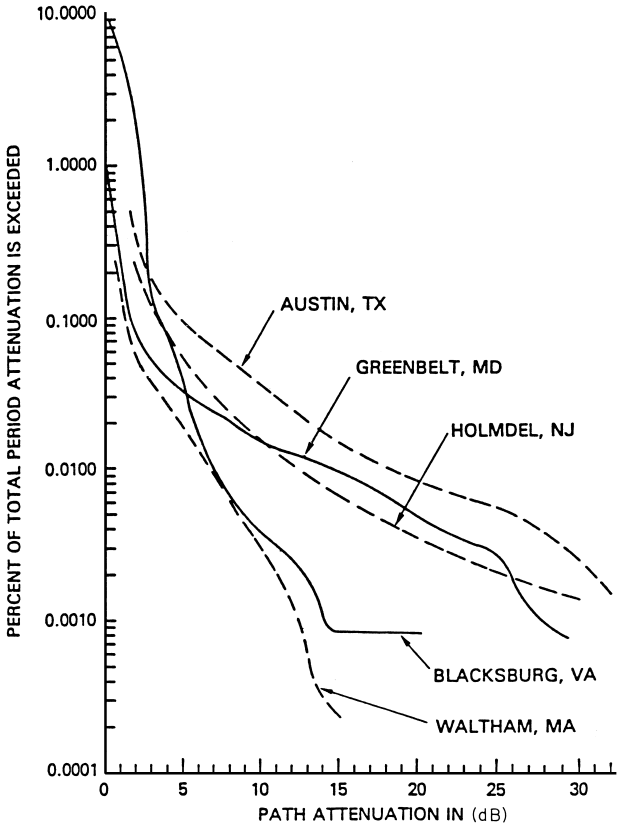


FIGURE 26 Comparison of down-annual 11.7 GHz attenuation distribution of measurements at five locations adjusted to 30′ elevation angles.

a given region. A typical attenuation probability distribution applicable to the United States is shown in Fig. 26 for K_u-band frequencies.

Also, increased noise temperature accompanies the path attenuation impairment. Rain raises the noise temperature of the system since it is at 290 K ambient, which is much higher than the background of space. The resulting noise temperature can be expressed in terms of the path attenuation a by

$$T_S = [(a-1)/a]T_O + T_{SO},$$

where T_S is the system noise temperature with rain, T_{SO} is the clear sky system noise temperature, and T_O is the ambient temperature, which is nominally 290 K.

The attenuation a is related to the rain attenuation L expressed in decibels by the expression

$$L = 10\log(a).$$

Thus, for example, if the attenuation due to rain is 6 dB, the value of a is 4 and the resulting system noise temperature becomes $3T_O/4 + T_{SO}$. For a typical clear sky system noise temperature of 200 K, the rain system noise temperature is 417.5 K. The noise temperature increase is not apparent on the up-link because the satellite is looking at the earth, which is of course always at an ambient temperature of 290 K. Thus, in calculating link budgets it is necessary to take into account both the path loss and the noise temperature increase on the down-link but only the path loss on the up-link.

Another form of atmospheric impairment that must be accounted for is that due to depolarization. Most satellite systems in service today use two senses of polarization—either horizontal and vertical or right- and left-hand circular—to achieve frequency reuse of the assigned frequency spectrum. Antenna systems, using modern feed technology, are able to discriminate between these polarizations to the extent that the unwanted polarization is attenuated by 30 dB or more relative to the wanted polarization under clear sky conditions. However, under heavy rain conditions this discrimination can drop to as little as 15 dB, and this must be accounted for in calculating the link availability. This is accomplished by including an appropriate carrier to interference ratio component in the calculation of the link budgets applicable to the system under study.

B. Modem Impairments

Modem impairments are those caused by inaccuracies and imperfections of the modulation process to generate the ideal transmission signal for the applied baseband signal and the demodulation process to perfectly extract the baseband signal at the receiver. For simple AM and FM, the modulation and demodulation processes depend on linear arithmetic relations that are usually easy to realize in today's technology, and near perfect modulation and demodulation implementation can be expected. Any impairments that do exist are principally due to nonlinearity in realizing the functional implementation. Only in the case of extremely wide bandwidth FM, where imperfection may be encountered in deviating the frequency of the carrier at high baseband frequencies, may some impairment be encountered in the modulation process. In digital modulation, similar constraints are at play. Digital modulation can be accomplished by methods too numerous to discuss in detail in this article. The most popular are FSK, PSK, and QAM. Impairments in their implementation are experienced in the accuracy with which values of frequency, phase, and amplitude can be maintained through the transmission system.

C. Filter Impairments

Impairments to achieving accurate transmission are induced in the filters that the information bearing signals

must pass. These filters are needed to minimize the interference from other signals that share the transmission facility and noise from natural origins. The filters themselves can produce impairment if they cause disorder in the reconstruction of the wanted signal. A typical filter characteristic is illustrated in Fig. 27.

The shape of the filter in terms of its amplitude and phase response has been selected to reproduce the signal at the receiver with minimum contamination from noise and interference. The shapes of the desired amplitude and phase responses are bounded by tolerance margins within which deviations are allowed for the manufacturing process and system operation of the filter. These margins are determined by both analytic and experimental methods to produce an acceptable level of impairment in transmission of the wanted signal. Figure 28 illustrates a two-dimensional scatter diagram for the near ideal

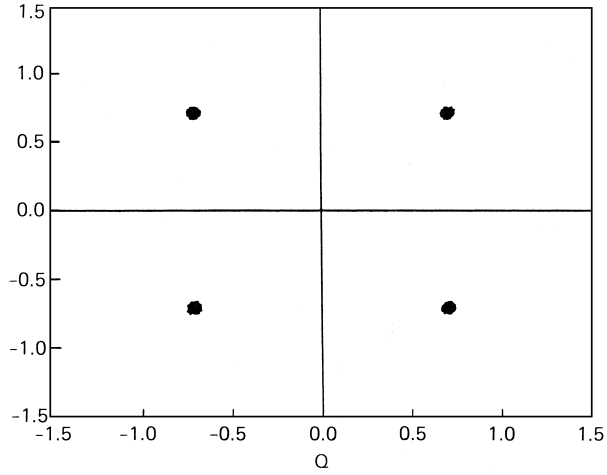


FIGURE 28 Two-dimensional scatter diagram: ideal transmission.

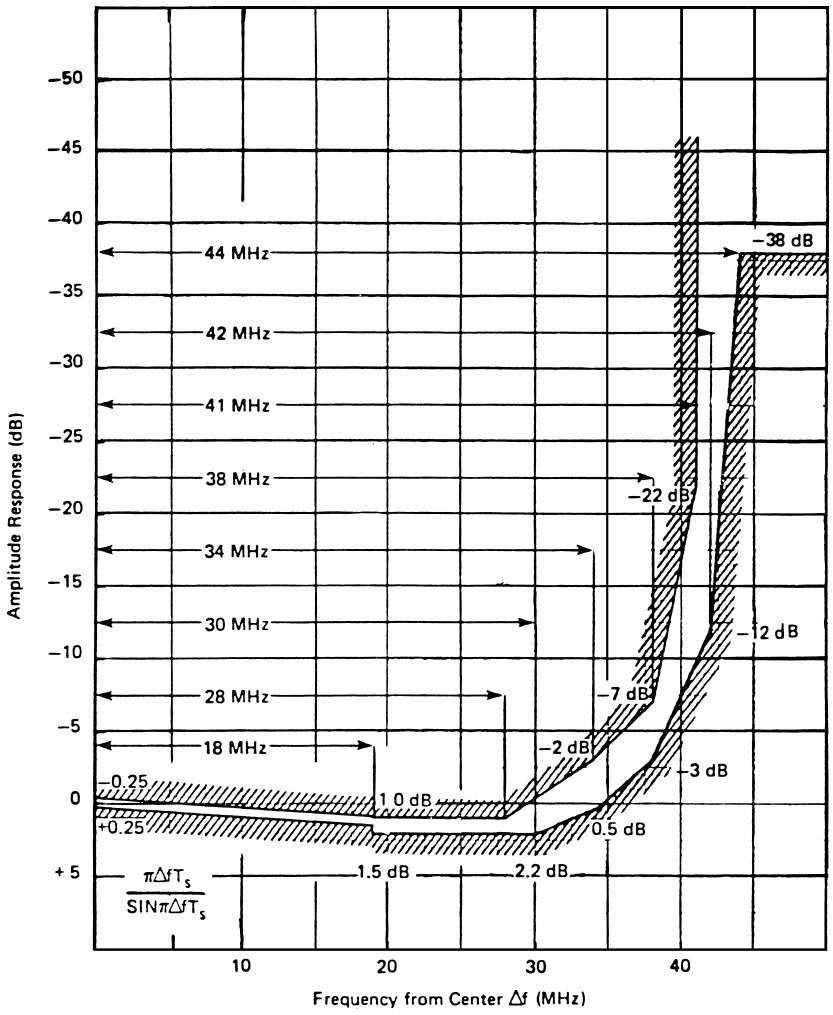


FIGURE 27 Modular filter amplitude mask for 120 Mbit/sec QPSK.

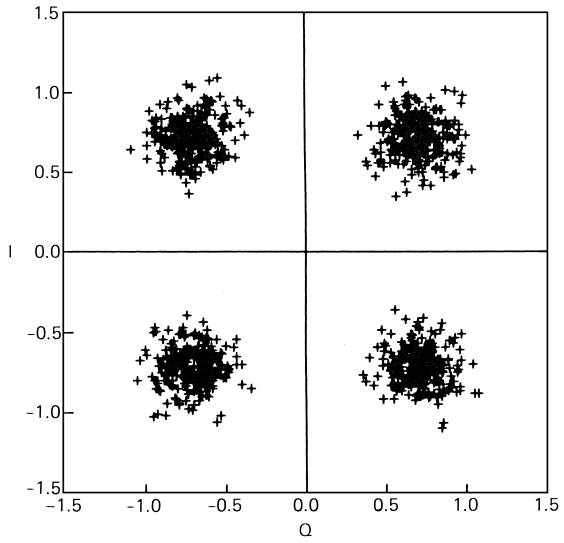


FIGURE 29 Two-dimensional scatter diagram: impaired transmission.

transmission of QPSK digital transmission. Each of the states of the transmission are clear and distinct, indicating that data transmission would be accomplished with high accuracy and hence low error rate.

Figure 29 is the same diagram but distorted due to improper filtering. The target areas are now dispersed and blurred, and the ability to select a particular state accurately is impaired and the error rate significantly increased.

Figure 30 shows the performance of QPSK modulation in terms of bit error rate.

D. Power Amplifier Impairments

In addition to impairments due to implementation, impairments can also be caused by imperfections in the transmission path. These are caused by a number of factors. The transmission path includes power amplifiers, which are inherently nonlinear. For example, if the satellite incorporates a traveling wave tube amplifier (TWTA),

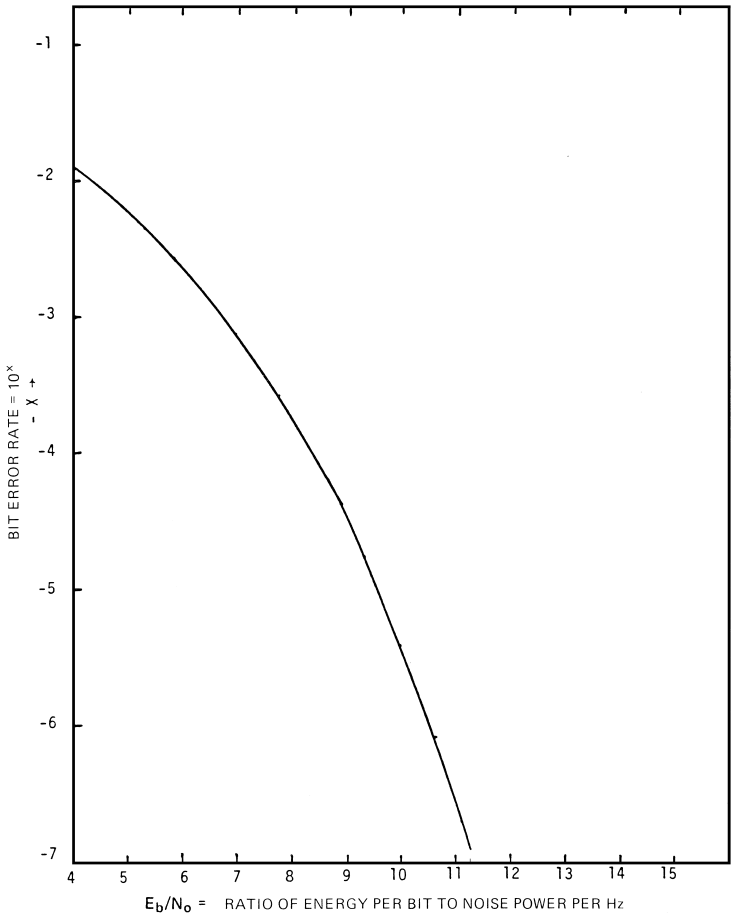


FIGURE 30 Theoretical performance of QPSK modulation in terms of bit error rate versus E_b/N_0.

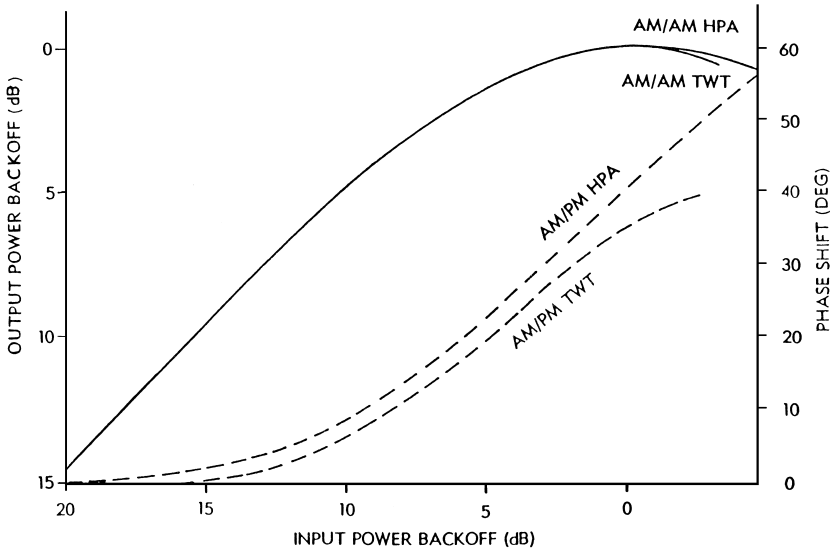


FIGURE 31 Nonlinear response curves for HPA and TWTA of INTELSAT V.

amplitude distortion, in the form of AM to AM, and phase distortion, in the form of AM to PM will be introduced. Typical TWTA amplitude and phase transmission characteristics are illustrated in Fig. 31.

The AM-to-AM and AM-to-PM effects cause the components of the transmission to interact, resulting in the generation of in-band intermodulation products. When the transmission is composed of many independent signals, as is the case of FDMA transmission, the intermodulation products create a random signal that can be treated like another source of random noise. However, when the transmission is a single signal, as is the case for single-carrier-per-transponder TDMA, it appears as a distortion of the signal wave shape that is dependent on the history of the signal, which when averaged over all possible states of the signal can be equated to an impairment equivalent to additive random noise or to an equivalent degradation in the performance of the digital demodulator. In general, to maintain the impairment resulting from TWTA nonlinearity to the equivalent of a carrier-to-noise ratio of 20 dB for FDMA transmission, it is necessary to operate the TWTA such that its output power is 4 dB less than its saturated power output. However, for single-carrier-per-transponder operation such as that used for TDMA, the impairment can be much less, depending on the bandwidth used to confine the channel. For digital transmission, if the bandwidth is equal to that of the symbol rate of the modulated signal, the TWTA can be operated to within 1 dB of its saturated power output. If the bandwidth is approximately 50% greater than the symbol rate, the TWTA can actually be operated beyond saturation. Figure 32 shows the performance of QPSK digital modulation with error correction coding in terms of the BER versus the ratio energy per bit to noise per unit bandwidth E_b/N_o, as experienced over a typical satellite transmission link.

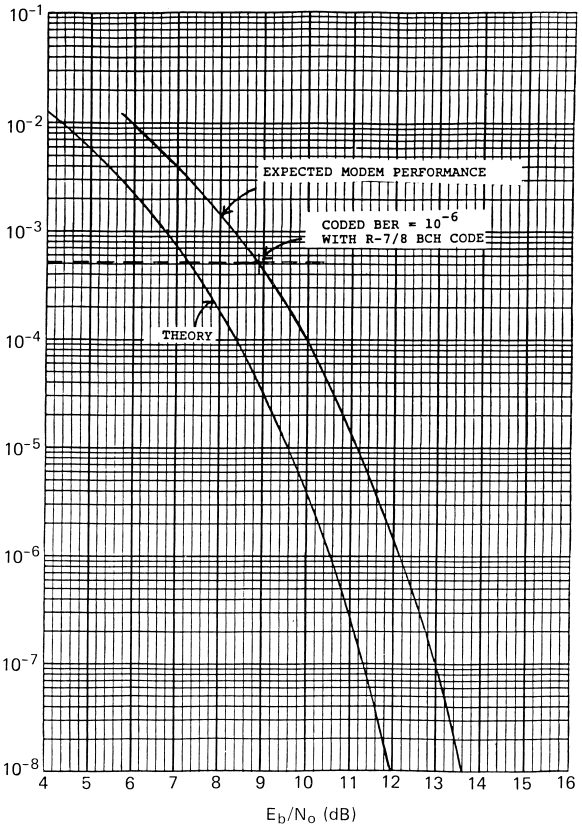


FIGURE 32 Modem performance over typical nonlinear satellite links.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SATELLITE SYSTEMS • COMMUNICATION SYSTEMS, CIVILIAN • GEOMAGNETISM • MODULATION • NETWORKS FOR DATA COMMUNICATION • RADIO PROPAGATION • SENSORS FOR CONTROL • SIGNAL PROCESSING, DIGITAL

BIBLIOGRAPHY

Campanella, S. J. (1981). *In* "Digital Communications-Satellite/Earth Station Engineering" (K. Feher, ed.), pp. 336–404, Prentice-Hall, Englewood Cliffs, NJ.

Campanella, S. J. (1985). "Satellite communications-challenges of the future," *Technical Marketing Society of America, Los Angeles, California. July 1985*, pp. 1–24.

Campanella, S. J. (1986). "Communications satellites of the future," *Am. Astronaut Soc., 33rd Annu. Meet., Boulder, Colorado, October 26–29, 1986*, pp. 1–21.

Campanella, S. J., and Harrington, J. V. (1984). "Satellite communications networks," *Proc. IEEE*. **72**(11), 1506–1519.

Campanella, S. J., and Pontano, B. (1986). "Advantages of TDMA and satellite-switched TDMA in INTELSAT V and VI," *COMSAT Tech. Rev.* **16**(1), 207–238.

Chang, P., Fang, R. J., and Jones, M. (1981). "Performance over cascaded and band-limited nonlinear satellite channels in the presence of interference," *ICC Conf., Denver, Colorado, June 1981*, pp. 47.3.1–47.3.10.

Edelson, B. 1. (1973). "Satellite communications for information networking," *J. Am. Soc. Inf. Sci.* **9,** 175–187.

Edelson, B. 1. (1977). "Global satellite communications," *Sci. Am.* **236**(2), 58–73.

Rider, L. (1985). "Analytic design of satellite constellations for zonal earth coverage using inclined circular orbits," *The Journal of Astronautical Sciences* **34**(1), 52.

ITU Radio Regulations, Vol. 1-E, 1998, pp. 34, 135.

NASA, Johnson Space Center, Orbital Debris Program Office.



Signal Processing, Acoustic

Edmund J. Sullivan
Naval Undersea Warfare Center

I. Basics
II. Signal Processing in the Time Domain
III. Spatial Processing and Antenna Arrays
IV. Noise and the Medium
V. Applications
VI. Recent Developments

GLOSSARY

A/D converter Device that samples an analog (A) quantity at some (usually) periodic rate and converts it to a digital (D) quantity of a specified number of bits.

Autocorrelation function The expected value of the product of a function of time at a given time t with its value at another time $t+\tau$. The time function (or time series) is said to be weak-sense stationary if the autocorrelation function depends only upon τ.

Biologics Biological sources of acoustic noise and scattering (usage in underwater acoustics and sonar).

Convolution The convolution of two functions is the integral of the product of two functions where one of the functions has its independent variable reversed in sign. The convolution of the two functions $f(x)$ and $g(x)$ is given by

$$c(y) = \int_0^y f(x)g(y-x)\,d\mathrm{x}$$

Decibel notation Logarithmic means of representing a quantity as compared with a reference quantity. For example, if $10\log(A/B) = +25\,(-25)$, then A is said to be 25 decibels (dB) "above" ("below") B.

Deconvolution Inverse process of convolution. If $c(y)$ is the convolution of $f(x)$ and $g(x)$, then the deconvolution would be the process that recovers $g(x)\,(f(x))$ given $f(x)\,(g(x))$ and $c(y)$.

Dirac delta function Usually designated $\delta(x)$, this function (more properly called a distribution) has the properties that it is zero everywhere except at $x=0$, and its integral over the range $-a$ to $+a$, where a is any number greater than zero, is equal to unity.

Doppler shift Shift in the frequency of a source, as perceived by the listener, that is caused by any relative motion along the line between the source and the listener.

Gaussian process: Any statistical process governed by the Gaussian probability density function. The Gaussian probability density function for a random variable x is given by

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2\sigma^2}$$

Here, σ is called the standard deviation of the process.

Gyre Large-scale closed circulatory system (oceanography usage).

Nyquist rate Lowest rate at which a finite-bandwidth signal can be periodically sampled in order to reproduce the signal completely and faithfully. The Nyquist rate is equal to twice the bandwidth of the signal.

Platform Physical location of some relevant component of a sonar system, such as an acoustic array. It commonly designates a ship or submarine.

Power spectral density The power in a one-Hertz band. It is given by the Fourier transform of the autocorrelation function.

Seismics Study of the structure of the earth by the use of sound.

Tomography: Process whereby a three-dimensional object is sectioned into plane sections for viewing. It is usually accomplished by performing an inverse scattering process using electromagnetic, acoustic, or X radiation.

White noise Noise whose spectrum is "flat," that is, constant as a function of frequency. It is characterized by an autocorrelation function that is a Dirac δ function.

ACOUSTICS is the science of sound, including its production, transmission, and reception. Sound itself is the propagation of mechanical energy in a material medium. As such, it can carry information. Acoustic signal processing is the discipline concerned with the separation of desired acoustic information from undesired information, or "noise."

I. BASICS

The transfer of acoustic information can be conveniently represented by a model with three parts: the source, the medium, and the receiver. The source transmits acoustic energy of a given form through the medium, and some of this energy, along with unwanted energy, or "noise," is picked up by the receiver. In practice, such systems can be found in either of two modes: the active or the passive. Examples of these two modes are realized in shipborne sonar (*so*und *na*vigation and *r*anging) systems. An active sonar system can be found in two forms. These are referred to as monostatic and bistatic and are depicted in Fig. 1A,B. The monostatic system has the source and receiver located on the same platform, and indeed both may employ the same acoustic transducer. The key point is that, for active systems, the source is under the control of the listener. In a passive sonar system, only the receiver is under the control of the listener, as in the case of a surveillance system that listens for sounds that may emanate from a submarine. This is shown in Fig. 1C.

These ideas can be made clearer by comparing sonar systems with radar (*ra*dio *d*etection *a*nd *r*anging) systems, which are always active. A radar system transmits electromagnetic energy and, as in sonar systems, receives scattered or reflected energy both monostatically and bistatically. Generally, there is no passive radar, however, since the "listener" must always provide the "illuminating" energy.

The comparison with radar systems can also be useful when discussing the medium. Both sonar and radar systems can suffer degraded performance due to scattering from matter in the medium and scattering from boundaries. The problem is somewhat more severe in the sonar system, however, since the boundaries are usually in closer proximity, and particulate matter and biologics in the medium are a major source of reverberation.

Another major difference between the two systems arises from the fact that sonar system performance is more profoundly affected than that of radar systems by the rather strong dependence of the speed of sound both on temperature and on pressure. Consequently, unlike the situation with radar systems, straight line propagation is essentially impossible in most cases in sonar systems. (In some acoustic applications other than sonar, such as seismics and tomography, the importance of the medium remains a major concern.)

Once the acoustic energy reaches the receiver, the signal-processing task begins. Acoustic signal processing is concerned with both detection and estimation. Detection is defined as the determination of the existence or nonexistence of a signal at the receiver. Estimation, which is a somewhat broader term than detection, is the quantification of certain parameters or descriptors of the signal, the medium, or the contents of the medium.

An example of detection is the determination of the presence of a submarine in a volume, termed the detection volume, by an active sonar system. That is, the only concern is the existence or nonexistence of an echo arising from the detection volume and due to a submarine. On the other hand, the determination of the existence of flaws in a piece of metal by the use of high-frequency sound would require more information from the signal than simply its existence. At the very minimum, an estimate of the time delay of the reflected acoustic pulse would be necessary.

The question of estimation leads to a class of acoustic signal-processing problems called inverse problems. This is not a well-defined concept and is best described in terms of its relation to the "forward" problem. The forward problem can be stated as follows. If the source transmits acoustic energy of a given form through the medium, what will the receiver receive? The inverse problem is concerned with the question: If the receiver receives a signal of a given form, what does this tell us about the medium, its contents, or the source?

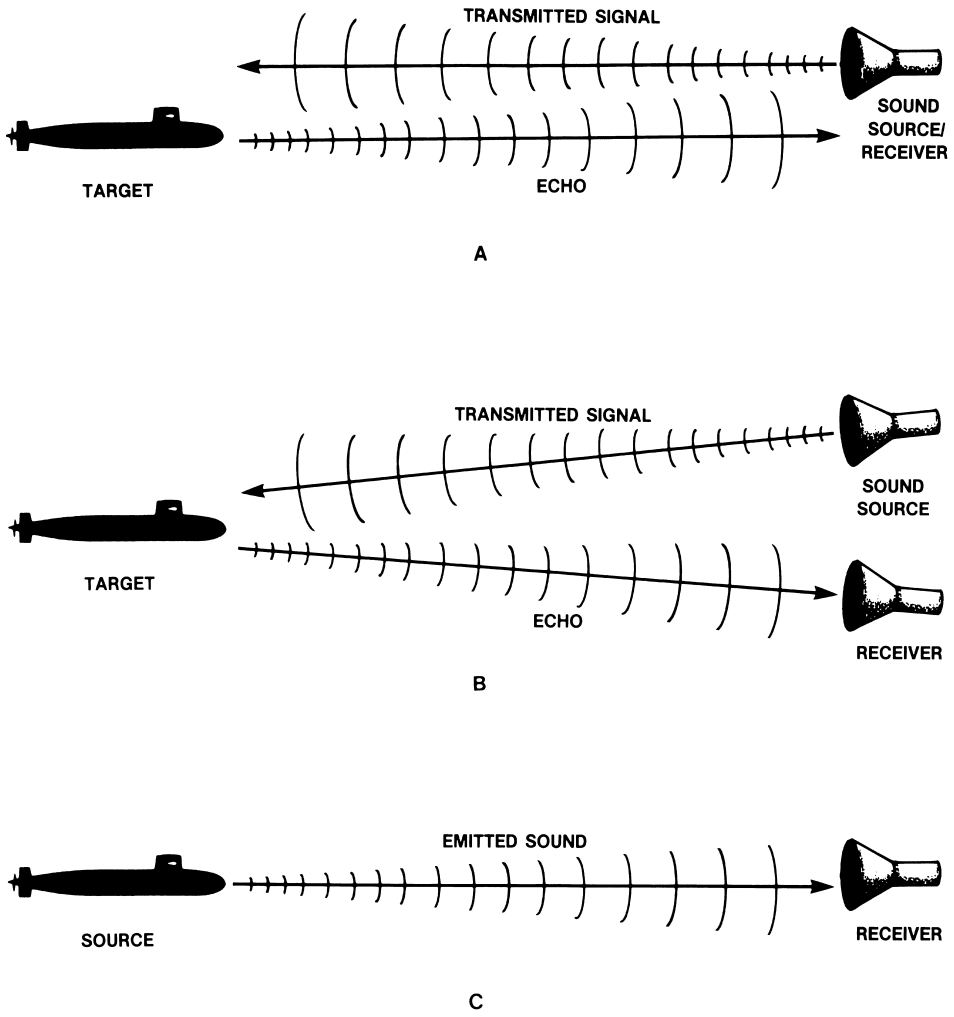


FIGURE 1 Sonar system configuration. A: active monostatic system; B: active bistatic system; C: passive system.

Examples of the forward problem are sonar, audio systems, and communication systems. The inverse problem is concerned with acoustic tomography, ocean bottom profiling, seismic deconvolution, and sonar target localization and classification. This perspective can be represented mathematically by Eq. (1).

$$f(\mathbf{x}) = \int K(\mathbf{x},\ \mathbf{x}')g(\mathbf{x}')\,d^3\mathbf{x}'. \qquad (1)$$

If Eq. (1) represented an acoustic scattering problem, then $K(\mathbf{x}, \mathbf{x}')$ would represent the acoustic energy scattered from a differential volume $d\mathbf{x}'$ or "source" at the point denoted by the vector coordinate $\mathbf{x}'$, and $g(\mathbf{x})$ would represent the distribution of sources and therefore the distribution of the scattering medium. The scattered field would be given by the function $f(\mathbf{x})$. The forward problem, then, would be concerned with determining $f(\mathbf{x})$ given some information about $g(\mathbf{x})$ and $K(\mathbf{x}, \mathbf{x}')$. An example of an inverse acoustic scattering problem would be the determination of $g(\mathbf{x})$ given some information about $K(\mathbf{x}, \mathbf{x}')$ and measured values of $f(\mathbf{x})$. Sometimes $K(\mathbf{x}, \mathbf{x}')$ is called the Green's function and here would describe the acoustic field scattered from a point "source."

Because the inverse problem is concerned with extracting a great deal more information from the signal than simply knowledge about its existence or nonexistence or the estimation of a few parameters, it can require a very computationally intensive process. In most cases the degree of computation is so great that only with the advent of modern computer technology has such processing become feasible.

Acoustic signal processing can be two-dimensional in the sense that it involves both spatial and temporal aspects. This is in contrast to certain communications problems, for example, that are concerned only with the time and frequency domains. The temporal aspect of the problem concerns itself with the behavior in time of the signal or its spectral characteristics. The spatial aspect concerns itself

mainly with the directional properties of the signal. This leads to the study of arrays or antennas whose spatial extent and directional sensitivities are of interest.

Signal-processing engineers are, in many cases, concerned with design of optimal or suboptimal systems. This basically leads to the design of optimal or suboptimal filters—that is, devices that efficiently extract the desired information in the time or frequency domain. The receiving array or antenna can be thought of as a spatial filter. A functional example of a spatial filter is the two-element array formed by the human ears, which enables us to determine the direction of incoming sound.

Acoustic signal processing is a methodology leading to many applications. These include audio systems, communications, acoustic holography, acoustic pattern recognition, acoustic tomography, seismics, ocean bottom profiling, nondestructive testing, speech recognition and synthesis, and sonar target localization and classification. In spite of the broad range of applications, acoustic signal processing lends itself to a rather well defined formal exposition, which is the subject of the next section.

II. SIGNAL PROCESSING IN THE TIME DOMAIN

A. Totally Known Signals

In the case of acoustic signal processing one is interested in at least two kinds of signals. These are deterministic signals that are corrupted by noise and random signals that are corrupted by noise. Examples of deterministic signals corrupted by noise are active sonar and nondestructive testing with high-frequency sound. In both cases, a known deterministic signal is transmitted and subsequently received after being acted on by the medium and its inclusions and corrupted by noise. An example of the second kind of signal, a random signal corrupted by noise, is the case of a passive sonar listening to the acoustic emissions from a submarine.

When considering such signals, there are two levels of knowledge that one may seek. These are commonly referred to as detection and estimation. In detection, as discussed in the introduction, one simply desires to know whether the signal is present at the receiver. In the case of estimation, one seeks to estimate one or more parameters of the signal. In the case of active sonar detection, one can determine the likelihood of the presence of an echo. Estimation then might be the estimation of the amplitude, frequency, or time of arrival of the echo. More complicated cases of estimation fall under the category of inverse problems.

We begin with the problem of detection of deterministic signals corrupted by noise. Formally, detection is cast as a problem in hypothesis testing. Basically, one forms two hypotheses:

$$H_0 \equiv \text{hypothesis that signal is not present (null hypothesis),} \tag{2a}$$

$$H_1 \equiv \text{hypothesis that signal is present.} \tag{2b}$$

In terms of the received signal x, the actual signal s, and the noise n, these two hypotheses can be expressed as

$$H_0;\ x = n \tag{3a}$$

$$H_1;\ x = s + n \tag{3b}$$

Expressing the probability of occurrence of x under the hypothesis that H_1 is true as P_1, and the probability of occurence of x under the hypothesis that H_0 is true as P_0, a decision is made based on the likelihood ratio Λ. That is, if

$$\Lambda = P_1/P_0 > \eta, \tag{4}$$

where η is a number to be determined, it is assumed that the signal is present. Clearly, this decision is strongly influenced by η. Decreasing η will increase the probability that the decision "signal present" will be made. However, the cost of doing so will increase the probability that such a decision is wrong. Such wrong decisions are referred to as "false alarms." The strategy is to choose η as low as possible without exceeding an acceptable false-alarm rate (FAR). The procedure, then, is to (1) determine acceptable FAR, (2) determine η based on FAR, (3) determine Λ, and (4) compare Λ with η.

The determination of Λ is the task of the processor. Such a processor, that is, one whose output is the likelihood ratio Λ, is referred to as an optimum processor. As a simple example of these ideas, we choose a case where the signal is a constant equal to unity and the noise is a zeromean Gaussian process with variance $= \sigma^2 = 2$. Thus,

$$p_0 = \frac{1}{\sqrt{4\pi}} \exp\left[-\frac{x^2}{4}\right], \tag{5}$$

$$p_1 = \frac{1}{\sqrt{4\pi}} \exp\left[-\frac{(x-1)^2}{4}\right]. \tag{6}$$

These probability densities are shown in Fig. 2.

The probability that the signal exceeds the threshold x_T is equal to the area labeled region 2. Thus, the probability of false alarm is given by

$$P_{FA} = \int_{x_T}^{\infty} \frac{1}{\sqrt{4\pi}} \exp\left[-\frac{x^2}{4}\right] dx. \tag{7}$$

Similarly, the probability of detection, given that the signal is present, is given by

$$P_0 = \int_{x_T}^{\infty} \frac{1}{\sqrt{4\pi}} \exp\left[-\frac{(x-1)^2}{4}\right] dx. \tag{8}$$

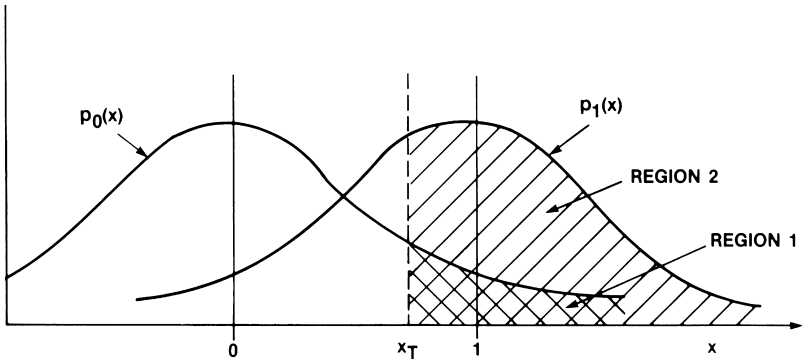


FIGURE 2 Probability densities for the two corresponding hypotheses. Here x_T is the threshold value. The area designated region 1 is equal to the false-alarm probability, and the area designated region 2 is equal to the probability of detection.

Choosing a false-alarm probability of 0.1, we can solve Eq. (7) for x_T. The result is $x_T = 1.8$. Since the likelihood ratio can also be written in terms of the probability density functions, η is given as

$$\eta = \frac{\exp[-(1.8-1)^2/4]}{\exp[-(1.8)^2/4]} = 1.9. \tag{9}$$

Thus, if η is picked to be 1.9, the probability of false alarm will be 0.1 and the probability of detection will be

$$P_D = \frac{1}{\sqrt{4\pi}} \int_{1.8}^{\infty} \exp\left[-\frac{(x-1)^2}{4}\right] dx = .285. \tag{10}$$

These results are valid for a single observation of x.

Although this is a simple example, the procedure to be followed in more complicated cases is, in principle, the same. The mathematics, however, becomes more complex.

On examination of Fig. 2, it can be seen that the greater the separation of the mean for signal present from that for signal absent, the higher the probability of detection can be for a given false-alarm rate. This separation is conveniently represented by the detection parameter d. For the example just given, d is defined as

$$d = (m_1 - m_0)^2 / \sigma^2, \tag{11}$$

where m_1 is the mean with signal present, m_0 the mean with signal absent, and σ the standard deviation of the noise. The quantities P_D, P_{FA}, and d are commonly combined and represented by a family of curves called receiver operating characteristics, or ROC curves, an example of which is shown in Fig. 3.

The preceding example is a binary decision problem; that is, there are only two hypotheses to be considered. The example is based on a single sample. In the case of time-varying signals, however, there are potentially an infinite number of samples available. It is possible, however, to reduce this case to a binary decision problem. An example of this is shown in the sequel.

Suppose one were to make N observations (samples) of a continuous time-varying signal. If these observations were independent samples of a Gaussian process, the joint probability densities for the two hypotheses H_0 and H_1 would be given by the products of the densities for each sample. That is,

$$p_0 = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x_i^2}{2\sigma^2}\right], \tag{12}$$

$$p_1 = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - s_i)^2}{2\sigma^2}\right], \tag{13}$$

where x_i and s_i are the ith measurement and signal samples, respectively, and $\Pi_{i=1}^{N}$ denotes multiplication of all N densities. The likelihood ratio is then given by

$$\Lambda = \frac{\prod_{i=1}^{N} \exp\left[-\frac{(x_i - s_i)^2}{2\sigma^2}\right]}{\prod_{i=1}^{N} \exp\left[-\frac{x_i^2}{2\sigma^2}\right]} > \eta, \tag{14}$$

where the $1/\sigma\sqrt{2\pi}$ term has been canceled. This relation can be conveniently expressed in logarithmic form. Equation (14) then can be expressed as

$$\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i s_i > \frac{1}{2\sigma^2} \sum_{i=1}^{N} s_i^2 + \ln \eta. \tag{15}$$

Making the assumption that the samples are taken at a frequency equal to twice the bandwidth of the noise (Nyquist sample frequency) and further assuming that the noise is white, that is, its power spectral density is equal to a constant, $N_0/2$, for all frequencies, the sums in Eq. (15) can be well approximated by integrals for the case of large bandwidth. One then finds

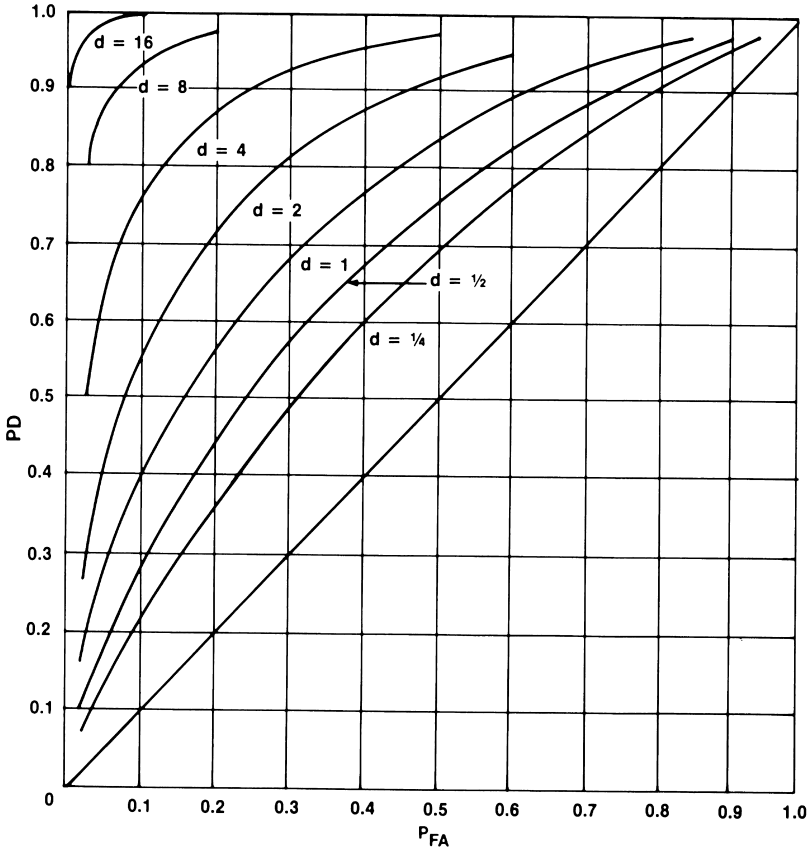


FIGURE 3 Receiver operating characteristics. The probability of detection is plotted as a function of the probability of false alarm for several values of the detection parameter *d*. [From Urick, R. J. (1967). "Principles of Underwater Sound for Engineers," McGraw-Hill, New York.]

$$\frac{2}{N_0}\int_0^T x(t)s(t)\,dt > \frac{E}{N_0} + \ln\eta. \tag{16}$$

Here,

$$E = \int_0^T s^2(t)\,dt \tag{17}$$

is the energy of the signal, T is the duration of the signal, and $N_0/2$ is the power spectral density of the noise, which is equal to the noise power in a 1-Hz band. In arriving at Eq. (16) we have used the fact that, for the case of Nyquist sampling, $\sigma^2 = (N_0/2) \times$ sample frequency.

The integral of the product of $x(t)$ and $s(t)$ is simply the correlation of the received signal with the known signal. This is to be compared with a threshold that is determined by η (which in turn is determined by P_{FA}, the false-alarm probability) and the signal-to-noise ratio. The point here is that, in spite of the fact that the observation vector can be very large, the decision statistic is a single number, the sufficient statistic. The detection parameter d now becomes

$$d = 2E/N_0. \tag{18}$$

The quantity $2E/N_0$ is the signal-to-noise ratio. Here, it is to be noted that the decision is based on only the sum (an integral) of the observations, weighted by a replica of the known signal. This is the sufficient statistic. Also, the ROC curves, shown in Fig. 3, now hold for this case with d as given above, where d depends only on the signal energy and the noise power spectral density.

The correlation process in Eq. (16) when directly implemented is referred to as a replica correlator. Its realization as a filtering process is expressed by setting

$$s(t) = h(T - t), \tag{19}$$

where $h(t)$ is the time domain representation of the matched filter. That is, the optimium (matched) filter is simply the time-reversed version of the replica and the process can be seen to be equivalent to a convolution in the time domain. The integration is carried out to time $t = T$, and the comparison for decision is then made. This is the optimum processor for totally known signals.

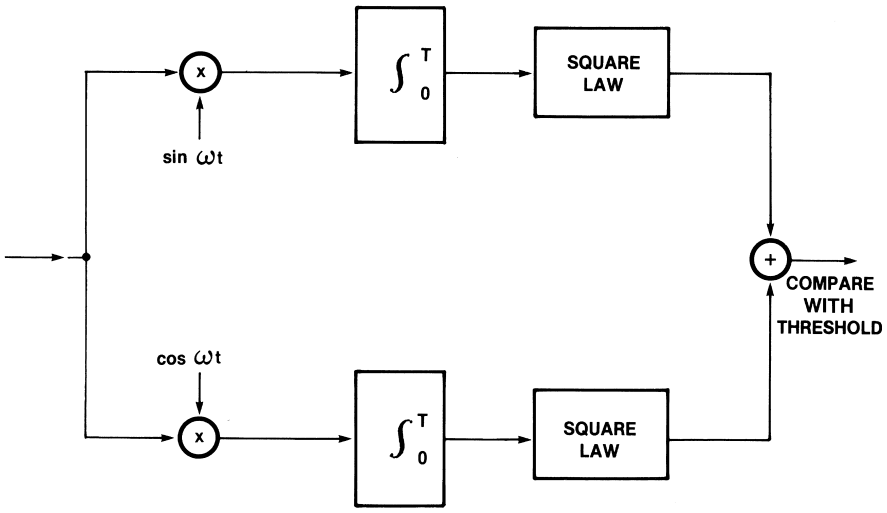


FIGURE 4 Quadrature detector. This constitutes the optimum linear processor for the case of a signal of unknown phase. [From Whalen, A. D. (1971), "Detection of Signals in Noise," Academic Press, New York.]

B. Not Totally Known Signals

The matched filter also plays an important role in the case of not totally known signals. In the case of unknown phase, the optimum detector turns out to be a square-law detector. This is usually realized in terms of a quadrature receiver, shown in Fig. 4. Here, it can be seen that the output is simply the sum of the squares of the outputs of two matched filters, where the signals being matched are $\sin \omega t$ and $\cos \omega t$, where $\omega = 2\pi f$ and f is the frequency of the signal. Clearly, it can be seen that the phase information in the signal is discarded, as one would intuitively expect. This can be viewed as an "incoherent matched filter."

Further cases, although more complicated, all involve either the coherent or the incoherent matched filter in their implementation. For the case of completely unknown signals, Gaussian noise, low values of the signal-to-noise ratio, and large time–bandwidth products, the detection parameter is

$$d = BT(S/N)^2, \tag{20}$$

where B is bandwidth, T integration time, S signal power, and N noise power. The term BT is usually referred to as the time–bandwidth product and occurs frequently in signal-processing formalism.

C. Estimation

So far, only the case of pure detection has been considered. When one is concerned with a determination of one or more of the signal parameters, the situation is conventionally referred to as a problem in estimation. In some cases of detection, an estimation of one or more of the signal parameters is obtained as a fringe benefit, but the original goal was purely that of detection. Here we are concerned with the direct estimation of one or more of the signal parameters as the actual goal. The subject of estimation covers a large area. For the purposes of the subject discussed here, only the so-called maximum likelihood estimate will be discussed, since it is the one most pertinent to the subject. The maximum likelihood estimate is based on finding the maximum of the likelihood function. This is a conditional probability density function that can be expressed as $p(x/a)$, where $p(x/a)\,dx$ is the probability that x lies in the interval dx given the parameter a.

The acoustic maximum likelihood problem is posed as follows. Given a received signal

$$x(t) = S(t, \alpha) + n(t), \tag{21}$$

it is desired to estimate the value of the parameter α.

Assuming that the noise is Gaussian and white, the likelihood function can be written

$$p(x/\alpha) = A \exp\left\{-\frac{1}{N_0}\int_0^T [x(t) - S(t, \alpha)]^2\, dt\right\}. \tag{22}$$

This is simply a continuous-time representation of Eq. (13). The quantity A is a normalization factor and is of no concern to us here. Upon maximizing this function with respect to the parameter α, the result is

$$\int_0^T [x(t) - S(t, \alpha)]\frac{\partial S(t, \alpha)}{\partial \alpha}\, dt = 0. \tag{23}$$

The estimation of the amplitude offers a simple example of how Eq. (23) is used. If the signal is written as

$$S(t, \alpha) = \alpha S_0(t), \tag{24}$$

it quickly follows that the estimate is the solution of

$$\int_0^T [x(t) - \alpha' S_0(t)] S_0(t)\, dt = 0 \tag{25}$$

or

$$\alpha' = \int_0^T x(t) S_0(t)\, dt \Bigg/ \int_0^T S_0^2(t)\, dt. \tag{26}$$

Note that, for a given integration time T, the denominator can be considered to be a constant, and thus the amplitude estimator is equivalent to a matched filter. If $S_0(t)$ is picked such that

$$\int_0^T S_0^2(t) = 1, \tag{27}$$

then the mean of the estimate is given as

$$E\{\alpha'\} = E\left\{ \int_0^T [\alpha S_0(t) + n(t)] S_0(t)\, dt \right\} = \alpha, \tag{28}$$

since the mean of $n(t)$ is zero. Here, E indicates the expected value operation. When the expected value of the estimated parameter is equal to the true value, the estimator is said to be unbiased.

The preceding example was concerned with the estimate of a single parameter. The process can be generalized to the estimation of multiple parameters.

A further increase in the complexity of the estimation problem leads to a class of problems generally referred to as inverse problems. Examples of acoustic inverse problems are tomography, or three-dimensional viewing via "sectioning" of an object; acoustic imaging; ocean bottom profiling; sonar target localization and classification; and seismics.

There exists no clear definition of when a given problem is an inverse problem. As mentioned in the introduction, an inverse problem is generally deemed to be inverse to a forward problem. It usually is more computationally intensive than the forward problem. The general linear acoustic inverse problem is conveniently discussed in terms of Eq. (1), which is reproduced below in two forms.

$$f(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{x}') g(\mathbf{x}')\, d\mathbf{x}'. \tag{1a}$$

$$f_i = \sum_i K_{ij} g_j. \tag{1b}$$

For convenience, we assume that Eq. (1) represents the general linear acoustic scattering problem. Here $g(\mathbf{x})$ represents the "source" or scatterer distribution, $K(\mathbf{x}, \mathbf{x}')$ represents the scattering from a differential volume, $f(\mathbf{x})$ represents the scattered field, and $\mathbf{x}$, $\mathbf{x}'$ are position vectors.

The acoustic scattering problem leads to an equation of the form of Eq. (1a) as follows. Starting from the inhomogeneous Hemholtz equation, we have

$$(\nabla^2 + k^2)\phi(\mathbf{x}) = s(\mathbf{x}). \tag{29}$$

Here, $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ in Cartesian coordinates. This operation is known as Laplacian; k is the wave number and is given by $k = 2\pi/\lambda$, where λ is the wavelength. An integral equation solution for the scattered field $\phi_s(x)$ can be found as

$$\phi_s(\mathbf{x}) = \int s(\mathbf{x}') \frac{e^{ik|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x} - \mathbf{x}'|}\, d\mathbf{x}' \tag{30}$$

where $\phi(x)$ in Eq. (29) is the total acoustic field and is given by

$$\phi(\mathbf{x}) = \phi_i(\mathbf{x}) + \phi_s(\mathbf{x}) \tag{31}$$

and $\phi_i(\mathbf{x})$ is the incident field. Equation (30) simply states that the scattered field at the point denoted by the vector $\mathbf{x}$ is that component scattered from a differential volume $d\mathbf{x}'$ located at $\mathbf{x}'$ summed over the volume distribution $s(\mathbf{x})$. The problem, then, is to invert the system to obtain $s(\mathbf{x})$, which is a description of the distribution of the scatterers.

Because of its generality, the inverse problem does not lend itself to a formal structure in the same sense as do detection and estimation. It is probably more realistic to consider it a collection of techniques. There are, however, some insights that can be gleaned from examination of Eq. (1). First, it is clear that it is generally an integral equation problem. Except for rare special cases, a solution in closed form will not be possible, so that a numerical solution will be called for. If the data are in discrete form, it becomes a problem in linear equations. Also, there can arise severe mathematical problems in the solution of such problems. One of these is that the problem can be what mathematicians refer to as ill posed. This is a problem that can be due to noisy data, not enough data, or a combination of the two. A problem is well posed when a unique solution exists that is stable to changes in the data. The problem of nonuniqueness can, in many cases, be dealt with by introducing *a priori* information.

The stability problem arises when, given the existence of a solution, the solution is extremely sensitive to small perturbations in the data. As an example, consider a measurement of $f(x)$ denoted by $f(x_1)$. Let this measurement incur an error Δf, and let f_0 be the true value. Then,

$$f(x_1) = f_0 + \Delta f = \int K(x_1, x)[g(x) + \Delta g]\, dx, \tag{32a}$$

$$f_0 = \int K(x_1, x) g(x)\, dx. \tag{32b}$$

Subtracting Eq. (32b) from (32a) yields

$$\Delta f = \int K(x_1, x) \Delta g(x)\, dx. \tag{33}$$

It can be seen from Eq. (33) that Δf can be thought of as a weighted average of $\Delta g(x)$, where $K(x_1, x)$ is the weight. Thus, one is free to select a function $\Delta g(x)$ whose weighted average is as close to zero as desired, but can produce large errors on $g(x)$.

A heuristic explanation of how the scattering problem can be ill posed is to note that the scattered field, which is the source of the observed data, can depend rather weakly on rather large changes in the scattering object or objects. Thus, a small amount of noise or simply the addition of a new datum can represent a very large fictitious change in the scattering object, which is the object of interest.

A second problem that can arise is that an inordinate computational load may ensue. It is, as previously mentioned, this very problem that prevented many inverse problems from being solved until the advent of modern computational capabilities. As a result, much of the effort expended in the inverse problem is focused on particular formulations and algorithms.

III. SPATIAL PROCESSING AND ANTENNA ARRAYS

A. Acoustic Transducers

As previously mentioned, acoustic signal processing is a two-dimensional problem in that it involves both time domain processing and spatial processing. The spatial-processing aspect of the problem concerns itself with the directional aspects of the transmission and reception of the sound. This is accomplished by means of transducers. The term *transducer* is used in many ways. It sometimes means a single device; other times it is used to describe an array of devices. Historically, in the case of underwater sound, the transmitting transducer was called a projector and the receiving transducer or sensor was called a hydrophone. Here, the term *transducer* means a single device that is either a transmitter or a receiver, or sometimes both. A collection of transducers is called an array.

Most receiving transducers are usually configured such that their output is a voltage that is proportional to the acoustic pressure at the transducer. For this reason, the pressure is usually the variable of interest. When discussing the flow of acoustic energy, it is convenient to define the acoustic intensity. This is the average rate of flow of energy across a unit area, where the direction of flow is normal to the area. For the case of a fluid medium, the pressure and intensity of a plane wave are related by

$$I = \overline{\rho^2}/\rho c, \tag{34}$$

where $\overline{\rho^2}$ is the time average of squared pressure, ρ the density of the fluid medium, and c the speed of sound in the medium.

Excluding audio microphones, there are generally four types of transducers: electromagnetic, piezoelectric, magnetostrictive, and optical. As suggested by its name, the electromagnetic transducer operates as a projector by utilizing the magnetic force produced by an electromagnet in whose coil an appropriately modulated electric current exists. An electromagnetic receiving transducer uses the principle of magnetic induction. The propagating acoustic pressure wave causes a motion of a permanent magnet in a coil. An electric current is then produced in the coil. A magnetostrictive transducer uses the fact that some materials will change their dimensions under the action of a changing magnetic field. This effect, in similarity to the electromagnetic effect, is reciprocal (i.e., straining the magnetostrictive material changes the magnetic field and thereby produces an electric current). Hence, there are magnetostrictive receiving transducers. There are several known magnetostrictive materials, some having higher saturation strain (maximum displacement) than others. Early alloys of nickel, iron, and cobalt provided modest saturation strains whereas more recently developed materials, which combine iron with various rare earth alloys, such as terbium, dysprosium, samarium, etc., permit much higher saturation strains. Also, magnetostrictive transducers are typically low frequency devices since there are intrinsic eddy current losses which counteract the applied magnetic field at higher frequencies. Piezoelectric transducers, which are also reciprocal, make use of the fact that certain materials change their dimensions under the action of an electric field. Some of these materials occur naturally, such as quartz or Rochelle salt, while a great number are synthesized chemically such as barium titanate, lead zirconate titanate (PZT), lead magnesium niobate (PMN), and lead zinc niobate (PZN).

For broadband operation, piezoelectric polymers such as polyvinylidene fluoride (PVDF), or piezocomposites (which combine the hard piezoceramic materials mentioned above with a soft viscoelastic polymer), are the materials of choice. Both the piezoelectric polymers and the piezocomposites are available in sheet form and may be patterned with a conductive material to form array elements. Also, they can be used on curved surfaces, such as ship or torpedo hulls, to form so-called conformal arrays. Optical transducers are not reciprocal. They operate only as receivers of sound. They generally fall into two categories: coherent and incoherent. Coherent optical transducers detect sound by using a change in the phase of very narrowband light energy produced by a laser. Although this effect can be produced in many ways, one way would be to use the acoustically induced motion of a mirror to change the optical path length of light compared with a reference beam of light whose path length is not affected by the sound. When the two beams are combined,

the sound is detected as a change in the optical interference pattern. Incoherent optical transducers depend on a change of light intensity caused by the sound for their operation. Again, this can be accomplished in many ways, one of which is to use a cladding on an optical fiber whose index of refraction is affected by the sound pressure. The cladding is a material surrounding the fiber that has an index of refraction less than that of the fiber and is therefore responsible for the capacity of the fiber to contain the light with little leakage. This type of transducer modulates the light in the fiber by acoustically controlling the amount of light that leaks out of the fiber.

B. Acoustic Arrays

By combining acoustic transducers in an array, one can control the directional properties of acoustic transmission and reception. For example, in many cases it is desirable to transmit sound only in a particular direction. Similarly, it may be important to receive sound only from a particular direction, to separate an acoustic signal from noise, or to resolve received sound into its various sources. In the case of reception (which is all that is considered here) such operations on the received sound waves are referred to as spatial processing and are accomplished by a combination of an array of transducers and a beam former.

Consider Fig. 5, where a plane wave of sound pressure, assumed to be at a single frequency, is arriving with a direction of propagation given by θ. The N point receiving transducers are equally spaced with spacing d in a line whose length is given by $L = d(N-1)$. The quantity L is commonly referred to as the aperture of the array. The output of each transducer is assumed to be a voltage whose value is proportional to the instantaneous acoustic signal, where the acoustic signal at the hydrophone labeled $n = 0$ is taken to be the real part of

$$A_0(t) = A_0 e^{i\omega t} = A_0[\cos(\omega t) + i\sin(\omega t)], \tag{35}$$

where $\omega = 2\pi f$ and f is the frequency of the sound. The nth transducer, then, has output voltage given by

$$V_n(t) = V_n \exp[i\omega(t + \tau_n)], \tag{36}$$

where it is easily shown that

$$\tau_n = nd\sin\theta/c. \tag{37}$$

Here, $c = f\lambda$ is the speed of sound and λ the wavelength.

Substitution of Eq. (37) into Eq. (36) yields

$$V_n(t) = V_n \exp(i\omega t + inkd\sin\theta), \tag{38}$$

where, as before, k is the wave number and is given by $k = 2\pi/\lambda$. The quantity k can be thought of as a spatial frequency. Since we are interested only in the spatial aspects of the problem, we shall set the time t equal to zero to simplify matters. In the case of the conventional beamformer, the outputs of the N transducers are now added together, resulting in

$$V(\theta) = \sum_{n=0}^{N-1} V_n \exp(inkd\sin\theta). \tag{39}$$

Equation (39) describes the spatial (angular) response function of a linear array of equally spaced transducer elements to a plane-wave acoustic signal impinging on it at angle θ. The term $nkd\sin\theta$ is the phase of the nth output with respect to the first transducer ($n = 0$).

The absolute magnitude of Eq. (39) is plotted in Fig. 6 for the case $d = \lambda/2$ and all V_n equal. This function is

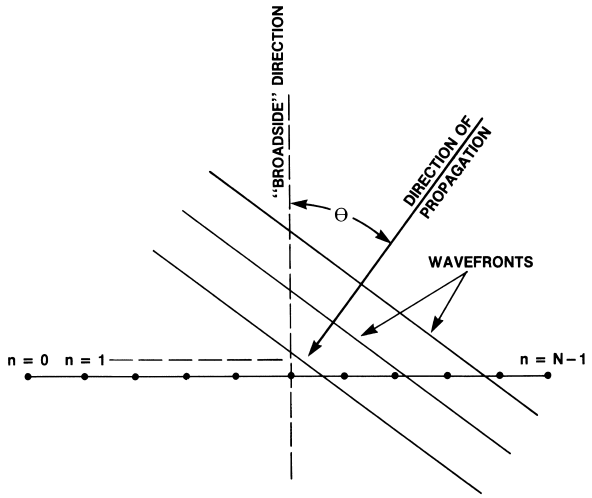


FIGURE 5 Line array of uniformly spaced point elements. The plane waves are impinging on the array at an angle θ from the 0°, or "broadside," direction.

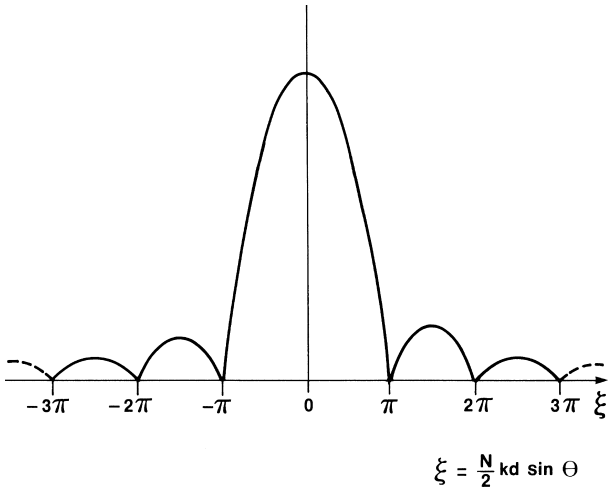


FIGURE 6 Response function of a line array of equally spaced point elements.

approximately equal to $|\sin(\xi)/\xi|$. As the spacing d approaches zero and the number of elements approaches infinity, the approximation approaches an identity.

A great deal can be learned about the properties of linear arrays by inspection of Eq. (39) and Fig. 6. First, the array has a region of maximum response to signals centered at $\theta = 0^\circ$ (broadside). This region of maximum response is usually referred to as the main lobe, and the other maxima are called sidelobes. Second, the response drops off rapidly as the signal is moved away from broadside. The power response is down to about one-half of the maximum when

$$\frac{N}{2} kd \sin\theta' = \frac{\pi}{2}. \tag{40}$$

$\Delta\theta = 2\theta'$ is referred to as the half-power beamwidth.

Solving for this beamwidth from Eq. (40) results in

$$\Delta\theta \simeq \lambda/Nd, \tag{41}$$

where the approximation $\sin\theta \simeq \theta$ for small θ has been used. Since $Nd \simeq L$, for large N, it follows from Eq. (41) that

$$\Delta\theta L' \approx 1, \tag{42}$$

where $L' = L/\lambda$ is the aperture in units of wavelength. This is sometimes referred to as the beamwidth–aperture product and shows that, to achieve a given beamwidth, the nondimensional aperture must be approximately equal to the reciprocal of the desired beamwidth in radians. Finally, by inspection of Eq. (39), it is seen that $V(\theta)$ is a periodic function that repeats at values of θ that satisfy

$$\sin\theta = \pm q \frac{\lambda}{d}, \tag{43}$$

where q is an integer. Thus, if d/λ is equal to unity, then the $V(\theta)$ has other maxima at $\theta = \pm 90^\circ$ (end fire). For larger values of d/λ, more maxima occur. This phenomenon is called aliasing and is a consequence of the aperture being undersampled. In practice, it would cause an ambiguity in the determination of the arrival angle of a signal. (These extra main lobes are sometimes called grating lobes.) For this reason, array elements are usually spaced at $d = \lambda/2$ or less. This is the spatial version of Nyquist sampling.

On further inspection of Eq. (39), it is seen that the maximum response occurs where the relative phases between the element outputs is zero. Thus, the main lobe can be "steered" to any desired angle θ_s by introducing appropriate phase shifts δ_n, where

$$\delta_n = nkd \sin\theta_s \tag{44}$$

Thus, the expression for a line array of equally spaced elements, steered to angle θ_s, is

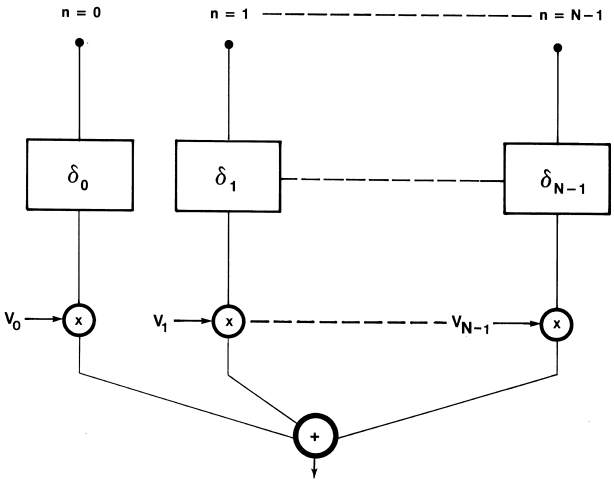


FIGURE 7 Phase-shift beam former. After the appropriate phase shifts are applied, the element outputs are multiplied by the weights, or "shading coefficients," before being summed.

$$V(\theta, \theta_s) = \sum_{n=0}^{N-1} V_n \exp(inkd \sin\theta - i\delta_n) \tag{45}$$

or

$$V(\theta, \theta_s) = \sum_{n=0}^{N-1} V_n \exp[inkd(\sin\theta - \sin\theta_s)]. \tag{46}$$

This is a simple example of beam forming and is illustrated in Fig. 7.

Rather than phase shifts, which have meaning only for the single-frequency case as treated here, the beam steering can be more generally effected by time delays, where

$$\tau_n = \frac{nkd \sin\theta_s}{2\pi f} = \frac{nd \sin\theta_s}{c}. \tag{47}$$

Equation (39) can be generally represented as an integral by defining an aperture function $V(x)$, namely,

$$V(\theta) = \int_0^L V(x) \exp(ikx \sin\theta)\, dx. \tag{48}$$

For the example of the uniformly spaced line array, the aperture function is

$$V(x) = \sum_{n=0}^{N-1} V_n \delta(x - nd), \tag{49}$$

where $\delta(x)$ is the Dirac delta function. On observation of Eq. (48), it is seen that $V(x)$ and $V(\theta)$ are Fourier transform pairs with transform variables $\kappa = k \sin\theta$ and x. That is

$$V(\kappa) = \int_0^L V(x) \exp(i\kappa x)\, dx. \tag{50}$$

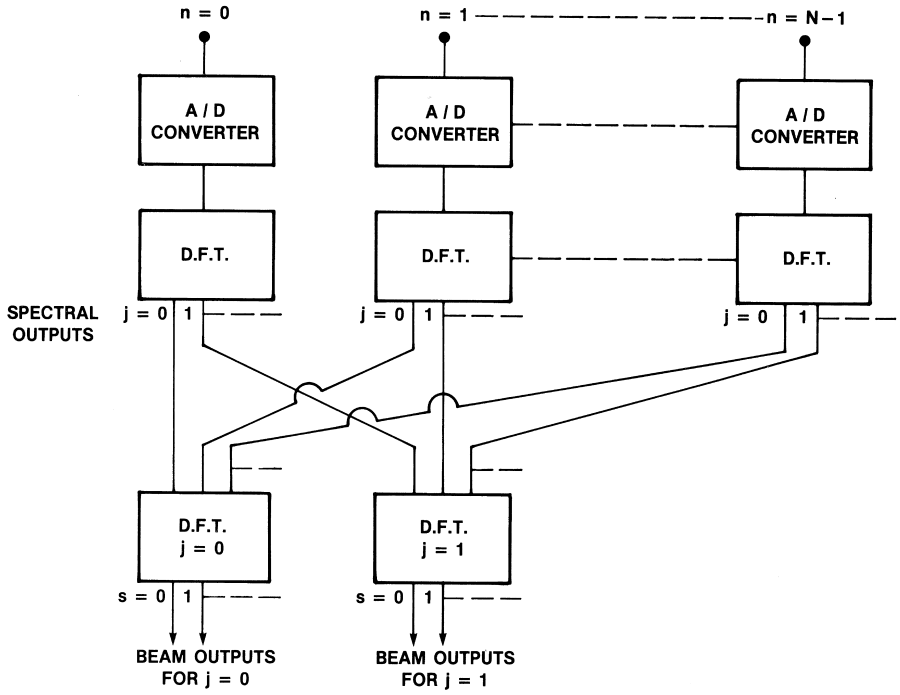


FIGURE 8 Frequency domain or $k-\omega$ beam former. The number of outputs is equal to the product of the number of beams and the number of spectral outputs (D.F.T.: discrete Fourier transform).

It is in this context that the acoustic array can be seen as a spatial filter, and the beamwidth aperture product naturally follows. Extending this analogy we rewrite Eq. (46) as

$$V(\theta, \theta_s) = \sum_{n=0}^{N-1} V_n' \exp(-inkd \sin\theta_s), \tag{51}$$

where V_n' represents the output of the nth element; that is,

$$V_n' = V_n \exp(inkd \sin\theta). \tag{52}$$

If we make the identification

$$\sin\theta_s = s\lambda/Nd, \tag{53}$$

where s is an integer, Eq. (51) becomes

$$V_s = \sum_{n=0}^{N-1} V_n' \exp(-i2\pi ns/N), \tag{54}$$

which is seen to be identical to a discrete Fourier transform of the element outputs, where the "spectral lines" are the beam outputs. For the case of broadband acoustic signals, Eq. (54) becomes

$$V_{sj} = \sum_{n=0}^{N-1} V_{nj}' \exp(-i2\pi ns/N). \tag{55}$$

Here, the index j labels the jth spectral output of a discrete spectrum analysis. This approach is referred to as $k-\omega$ beam forming, a possible realization of which is depicted in Fig. 8. Here, the element outputs undergo a discrete Fourier transform in the time domain to yield a set of narrowband signals, each of which is then Fourier transformed in the spatial domain to yield the beam set.

C. Array Optimization

Drawing again on the concept of the array as a spatial filter, the V_n in Eq. (39), sometimes referred to as shading coefficients, can be looked on as spatial filter coefficients. The beam-forming problem can now be stated in a more general form as the selection of the sets $\{V_n\}$ and $\{\delta_n\}$ to achieve an optimum performance, based on some set of criteria, against a given acoustic situation. There are two different approaches to this problem: pattern optimization and gain optimization. Pattern optimization proceeds first by the selection of a desired array response pattern that is deemed efficient for the given purpose and then, via some analytical or numerical technique, the selection of values of V_n and δ_n to best achieve this goal. A commonly desired pattern is one with a narrow main lobe and low sidelobes. This will give good performance against noise that is uniformly distributed in angle for an array that seeks to detect a signal arriving from a particular direction. A common example of this is the so-called Chebyshev shading, which gives the lowest possible sidelobe height for a given

beamwidth, given that all the sidelobes are constrained to be of the same height. Chebyshev shading is applicable only to a line array of uniformly spaced elements. When the array configuration becomes more complex, the behavior of the array response pattern is also more complex, such as in the case of arrays that are mounted on the curved surfaces of hulls or other platforms (usually called conformal arrays).

The gain optimization approach can be stated as follows. Given a particular noise field and signal, maximize the gain of the array, where the gain is defined as the signal-to-noise ratio of the array divided by the signal-to-noise ratio of a single element.

Suppose the noise amplitude is described by $\eta(\theta)$. Then, from Eq. (39), the total response of the array to this noise field is found by integrating over all directions, namely,

$$R_{\text{noise}} = \sum_{n=0}^{N-1} V_n \int d\Omega\, \eta(\theta) \exp(inkd \sin\theta), \qquad (56)$$

where $d\Omega$ represents the differential solid angle. The average noise power output is given by

$$P_{\text{noise}} = \sum_{n=0}^{N-1} V_n \sum_{m=0}^{N-1} V_m M_{n,m}, \qquad (57)$$

where

$$M_{n,m} = \int d\Omega\, \overline{|\eta(\theta)|^2} \exp[i(n-m)kd \sin\theta]. \qquad (58)$$

The overbar denotes the average.

Equation (58) is derived from Eq. (57) under the assumption that the noise field can be represented by spatially uncorrelated point sources. The optimization requires finding those values for the V_n that maximize the gain G, where

$$G = \frac{\text{signal power}}{\text{noise power}} = \frac{\sum_{n=0}^{N-1}\sum_{n=0}^{N-1} V_n V_m S_{n,m}}{\sum_{n=0}^{N-1}\sum_{n=0}^{N-1} V_n V_m M_{n,m}}. \qquad (59)$$

The quantity $S_{n,m}$ is the signal matrix and, for the case under consideration, is given by

$$S_{n,m} = \exp[i(n-m)kd \sin\theta_s].$$

Here, θ_s is the angle of the direction of the incoming plane-wave signal from broadside. Although the gain of an array is defined as the signal-to-noise ratio of the array divided by the signal-to-noise ratio of a single element, in this treatment, for simplicity, we assume that the signal-to-noise ratio at a single element is equal to unity so that G is also equal to the signal-to-noise ratio of the array as well as the gain.

In Eq. (59), $M_{n,m}$ is the covariance matrix of the noise, whose elements are proportional to the correlation of the noise outputs between the nth and mth elements. The conclusion to be drawn from Eq. (59) is that it is desirable to have uncorrelated noise and highly correlated signals. In the case where the noise is completely uncorrelated and the signal is perfectly correlated, the gain is

$$G = N. \qquad (60)$$

That is, the gain is equal to the number of elements in the array. This can be intuitively understood by noting that the signal amplitudes add coherently so that the signal power output is proportional to N^2, whereas the noise amplitudes add incoherently, giving a noise power output proportional to N. Equation (60) is usually written in decibel notation, that is,

$$G_{\text{dB}} = 10 \log N. \qquad (61)$$

When the noise is spatially uniform, the gain is called the directivity factor and, in decibels, it is called the directivity index.

D. Adaptive Arrays

Adaptive processing, which plays a role in most areas of signal processing, also finds application in array processing. Generally, adaptive processing uses information about the noise to maintain the optimality or near optimality of the system. As can be seen in Eq. (58), the covariance matrix is determined by $\eta(\theta)$, which is a description of the noise field. Since the optimal gain is determined by the covariance matrix, any changes in $\eta(\theta)$ would require a recomputation of the V_n. A sampling of the noise field is thereby required. In practice, this sampling could be effected by a set of auxiliary beams, either from the array itself or an auxiliary array. In some cases, a "suboptimal" approach is sufficient. In the case of a source of noise that is spatially localized, that is, one emanating from a single direction different than that of the signal, the output of the auxiliary beam that is receiving the noise can be subtracted from the output of the beam former. As intuitively expected, this would effectively produce a notch or "null" in the array response function. For this reason, adaptive processing against localized noise sources is sometimes called null steering.

IV. NOISE AND THE MEDIUM

A. Background

The medium can affect an acoustic-signal-processing problem in two ways. First, the propagation properties

of the medium may vary a great deal from place to place, or the medium may be strongly absorptive. Second, the medium may be a source of noise since contained matter or its boundaries can scatter the sound, or there may be sources of unwanted sound in the medium, such as biologics, shipping, wind, and rain. Also, in the high-frequency end of the spectrum, thermal noise, which is due to the thermal motion of the molecules, can be an important source of oceanic noise. Although the medium can be a problem in most acoustic-signal-processing problems, such as absorption of high-frequency sound by human flesh in the case of medical acoustic imaging, the case of the impact of the ocean on sonar problems will be emphasized here, since the ocean is a difficult—if not the most difficult—medium. As such, it has been extensively studied and can provide an excellent example of the problems encountered.

Figure 9 is a plot of the ambient noise in the ocean due to three sources: shipping, wind, and thermal sources. Here it is seen that which of the three sources is of importance depends on the frequency band of interest. Thus, in the case of most sonar problems, shipping and wind are important sources, whereas the thermal source becomes important only at relatively high frequencies. Two other sources of ambient noise in the ocean are rain and biologics.

A further source of noise, one that is important in the case of active sonar, is reverberation. As discussed, reverberation can occur due to scattering of sound from particulate matter and biologics in the volume of water. This is, appropriately enough, called volume reverberation. Reverberation from both the surface and bottom is called boundary reverberation. Reverberation can be particularly problematical since its level is proportional to the level of the transmitted signal. Thus, increasing the energy of the signal, which can give an increase in gain over ambient noise, is of no use in the case of reverberation. The level of the noise is given in terms of its spectrum level. This is the level in a 1-Hz band of the noise, which is considered to be broadband. A spectrum level of 10 dB would be expressed as

$$\text{spectrum level} = 10\,\text{dB re } 1\ \mu\text{Pa/Hz}. \tag{62}$$

This states that the level is 10 dB "above" a reference level of 1 μPa (micropascal) in a 1-Hz band. The micropascal is a metric unit of pressure equal to 10^{-6} N/m^2 (newton/meter^2). In the case of wind-related noise, there is a family of curves labeled sea state. This is a qualitative and somewhat subjective measure of the level of agitation of the sea surface. Sea state zero indicates a relatively calm sea, and at sea state 6, the ocean is quite agitated.

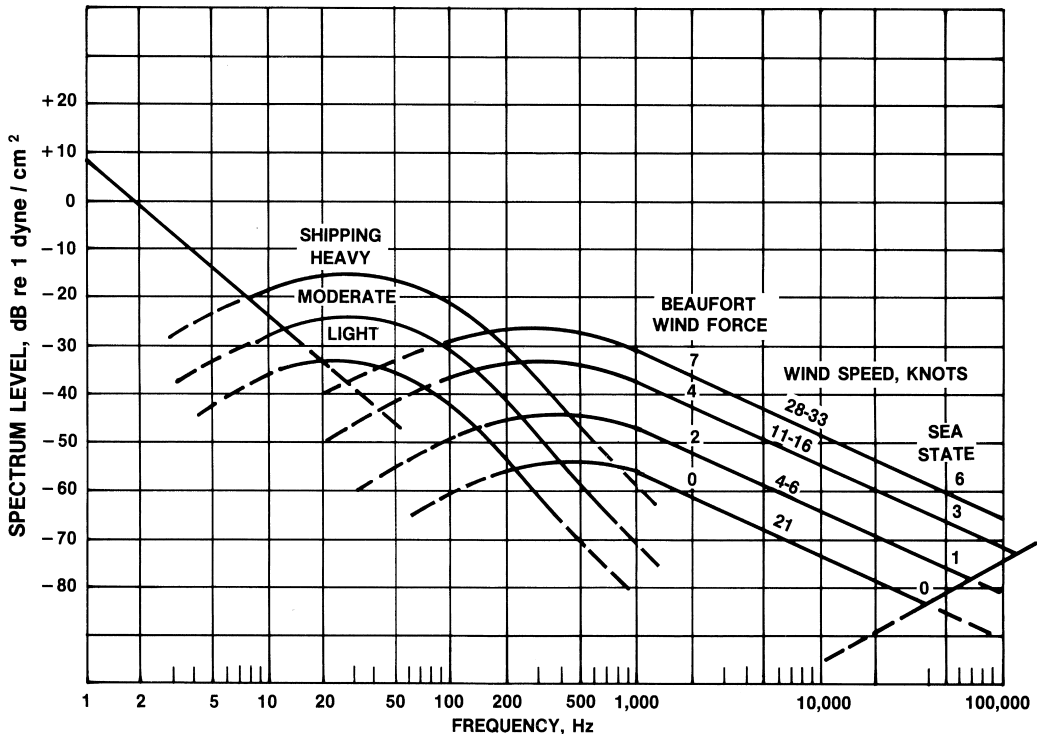


FIGURE 9 Ambient noise spectrum in the ocean. [From Urick, R. J. (1967). "Principles of Underwater Sound for Engineers," McGraw-Hill, New York.]

Next to its spectrum level there are two other properties of ambient noise that are important from a signal-processing standpoint. These are directionality and correlation. For the case of distant shipping, the noise is usually strongest within about $\pm 5^\circ$ of the horizontal, whereas the surface-generated noise would appear to be radiating downward. For the case of thermally generated noise, the directionality is generally isotropic. The spectrum of the noise is important both for the time-domain- and spatial-domain-processing problem, as can be understood from the preceding section. The spatial correlation and directionality play a strong role in the spatial processing tasks. As can be seen from Eq. (58), the directionality $\eta(\theta)$ strongly influences the covariance matrix, which in turn plays a major role in the determination of the "optimal" array.

The description of reverberation makes use of a quantity called scattering strength, where

$$s_{\mathrm{v,s}} = \frac{I_r/\text{unit volume}}{I_{\mathrm{i}}}. \tag{63}$$

Here, I_{i} is the intensity at the scattering volume, where the subscripts "s" and "v" refer to volume and surface, respectively. Theoretical expressions can be derived for I_{r} given the values of s_{v} and s_{s}, which can be computed theoretically for certain cases and, of course, can be measured. The volume reverberation level is then given by

$$\mathrm{RL} = 10 \log I_{\mathrm{s}} = \frac{s_{\mathrm{v}}}{r^4} I_0 A, \tag{64}$$

where I_{s} is the scattered intensity at the receiver, I_0 the intensity of the source, and r the distance to the scattering region; A is a number determined by the transmit and receive patterns of the sonar, the size of the scattering volume, and the characteristics of the sonar signal. This expression is not always useful, since the range to the reverberating volume may not be well defined, or the volume may, in fact, encompass the sonar system itself. A similar expression can be derived for surface reverberation.

Absorption can be a major problem in areas of acoustic signal processing other than sonar. In particular, acoustic imaging and nondestructive testing may involve media that produce excessive absorption. In the case of sonar, absorption is caused by certain salts contained in the ocean and is a function of the temperature, depth, and frequency. The major cause of this absorption is the acoustically driven disassociation of magnesium sulfate ($MgSO_4$). The effect of this phenomenon is shown in Fig. 10.

Other sources of noise can enter into an acoustic-signal-processing problem. Electrical noise—that is, noise arising from the electronics involved in the processing devices—can under certain circumstances cause problems. Also, a phenomenon commonly called flow noise can cause severe problems for receiving transducers and arrays on moving platforms. This noise arises from the fact that turbulent flow over the hydrophones can cause pressure fluctuations that appear as sound to the transducers. This can arise in the case of a sonar receiver mounted on the hull of a moving ship or submarine. Also, propulsion machinery can cause mechanical vibrations that can propagate through the structure of the vessel and introduce noise into the array.

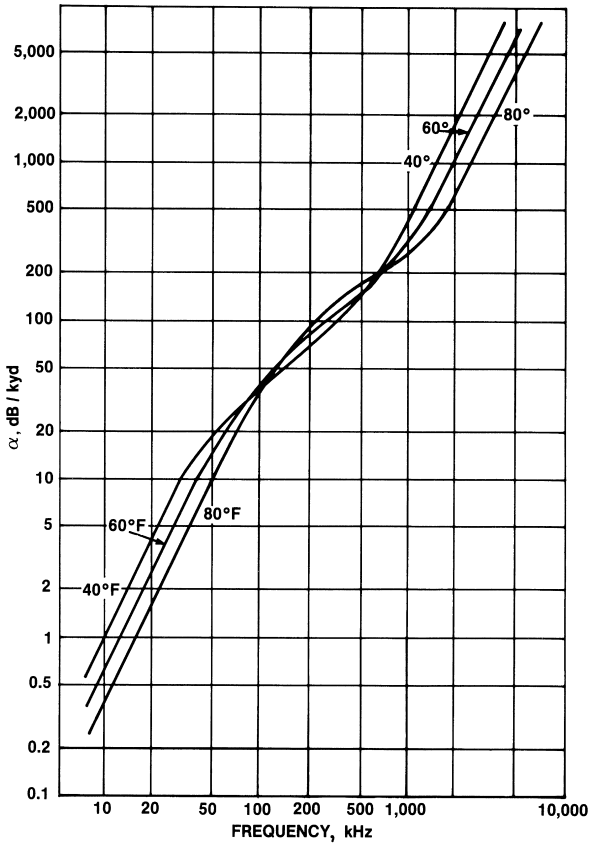


FIGURE 10 Sound absorption in the ocean as a function of frequency and temperature for a salinity of 35 parts per thousand. [From Urick, R. J. (1967). "Principles of Underwater Sound for Engineers," McGraw-Hill, New York.]

B. Reverberation Processing

Reverberation poses a major problem to sonar systems, and the means of processing against it are limited. One technique is to take advantage of any Doppler shifts in the frequency of the returned signal. If the sonar system is operating against a moving target, the echo will be Doppler shifted by an amount

$$\Delta f \simeq 2 \frac{v}{c} f_0, \tag{65}$$

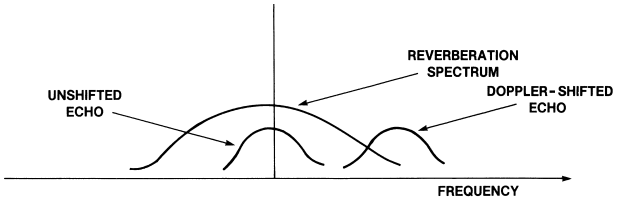


FIGURE 11 Spectra of Doppler-shifted and unshifted echo returns in reverberation.

where v is the speed of the target in the direction of the receiver and f_0 the frequency of the transmitted signal. Thus, if $2\Delta f$ is greater than the bandwidth of the reverberation, the spectrum of the echo will lie outside the spectrum of the reverberation and can easily be detected. This is illustrated in Fig. 11.

Note that the reverberation spectrum is broader than the echo spectrum. This is the case for volume reverberation and certain cases of surface reverberation. In surface reverberation, the spread is due mainly to surface motion caused by wind, whereas in volume reverberation the spread arises due to motion of the scattering particles themselves. Also, when the sonar source is moving, the Doppler shift of the reverberation depends on the direction from which it arrives at the receiver, thus contributing to the broadening of the spectrum.

Another means of processing against reverberation is to spread its spectral width somehow. If the spectrum of the reverberation can be spread, its power spectral density will be reduced since its power is proportional to the power spectral density times the bandwidth. One means of accomplishing this is with an FM slide or a pulse-modulated signal, both of which have increased signal bandwidth. An FM slide signal is a pulse whose frequency changes with time. This process is illustrated in Fig. 12.

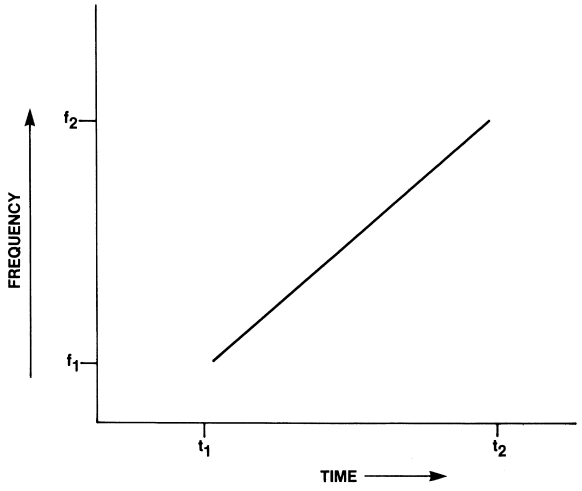


FIGURE 12 Linear FM slide.

If the frequency slide is not too rapid, the bandwidth of the pulse is approximately given by

$$\Delta f \approx f_2 - f_1. \tag{66}$$

In the case of a fixed-frequency pulse, the bandwidth of the pulse would be given approximately by

$$\Delta f \approx 1/T = 1/(t_2 - t_1). \tag{67}$$

Thus, an FM slide pulse can be designed whose bandwidth is greater than that of the fixed-frequency pulse of the same duration. When this pulse is processed by a matched filter (assuming that the target does not distort the pulse, an assumption that is not generally true), the detection parameter, and hence the signal-to-noise ratio, is increased over that for a fixed-frequency pulse of the same energy, since the power spectral density of the reverberation is reduced. This can be seen by referring to Eq. (18). Although the target does distort the echo somewhat, in many cases considerable gain against reverberation can still be obtained.

C. Effects of the Medium

The other problem addressed in this section is the impact of the medium on the processing problem. Again, the ocean provides a rather graphic example of how severe this can be.

The two major factors that have an impact on oceanic propagation are absorption and variations in the sound speed. The variation in sound speed is due mainly to pressure and temperature effects, as noted earlier. The speed of sound increases with increasing pressure and temperature. Thus, if the ocean were of a constant temperature, the sound speed would increase with depth. However, owing to solar and weather effects, the ocean is heated from its surface. This thermal layer, although subject to change with time and geographic location, has a thickness of the order of 1 or 2 km in deep ocean environments. It is subdivided into three sections. The surface layer is the part that is most strongly affected by short-term changes in the weather. It can undergo major changes in its temperature in a single day and can become quite mixed by the wind. This layer is of the order of 100 m in thickness. Beneath the surface layer lies the seasonal thermocline. This is a layer whose temperature gradient is stable for many days and undergoes major changes seasonally. This layer is also of the order of ~100 m. The next layer, which can extend to a depth of a few thousand meters, is called the main thermocline and possesses a temperature gradient that is stable over time scales of several years. Finally, below the main thermocline is the deep isothermal layer, where there is essentially no temperature gradient and the sound velocity dependence is determined primarily by the increase in

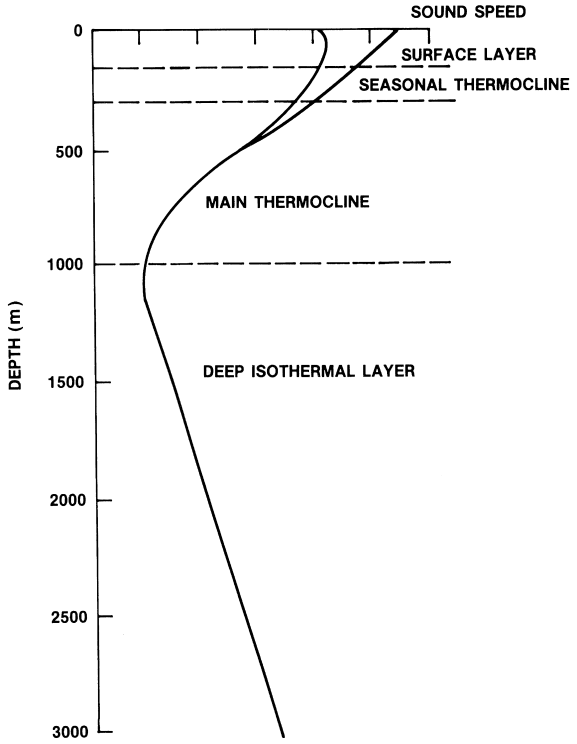


FIGURE 13 Typical oceanic speed-of-sound profile as a function of depth. [From Burdic, W. S. (1984). "Underwater Acoustic System Analysis," Prentice-Hall, Englewood Cliffs, NJ.]

pressure with depth. Figure 13 depicts this situation for a particular case. Here the sound speed is plotted against depth. An increase in sound speed with depth is called a positive gradient and will cause the direction of sound propagation to deflect upward. This can be intuitively seen by visualizing plane waves of sound propagating in a horizontal direction in the deep isothermal layer. Since the sound speed increases with depth, the lower regions of the plane waves travel faster than the upper parts, thus tilting them upward. This is depicted in Fig. 14. Thus, the ray (i.e., the line depicting the direction of propagation) curves upward. By the same argument, sound rays in a negative gradient will be deflected downward. This phenomenon can trap sound at the depth where the sound speed gradient is zero—that is, at the depth where the sound speed is minimum. This creates a duct called the deep sound channel, where sound can propagate for thousands of kilometers in deep ocean environments.

A second phenomenon is known as a surface duct. This occurs when the surface layer becomes mixed by the wind to the extent that there is no significant temperature gradient. The sound speed is then determined solely by pressure, and the layer then has a positive gradient. Since a positive gradient deflects (or refracts) the sound ray upward, the effect depicted in Fig. 15 ensues.

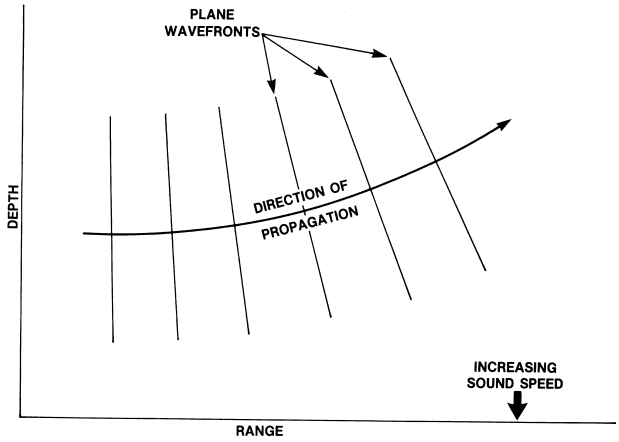


FIGURE 14 Bending or refraction of sound due to positive speed-of-sound gradient.

A third phenomenon is known as convergence zone propagation. This is also depicted in Fig. 15. The consequence of this effect is the existence of regions of increased intensity of the received sound at distances from the source that are multiples of ~50 km in deep ocean environments. These regions are separated by regions of very low sound reception.

Further complications arise when the ocean is rough, causing a scattering loss when sound is reflected from the surface, loss of energy when sound is reflected from the bottom, and shallow-water situations and propagation under polar ice, where there is essentially no surface layer.

As can be seen from the preceding, the study of the refractive and reflective phenomena of the ocean is

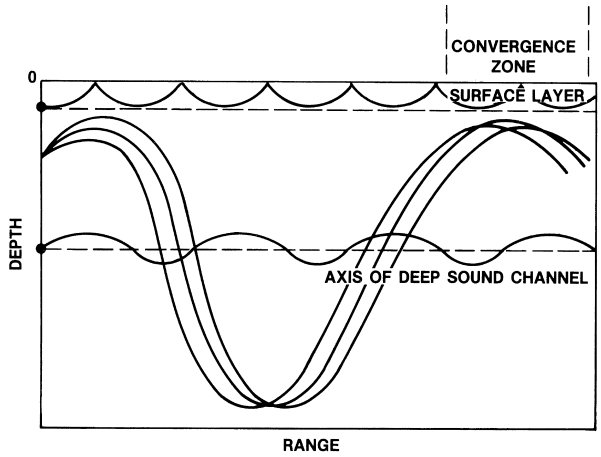


FIGURE 15 Effects of speed-of-sound profile on propagation. When the surface layer is mixed, the surface duct phenomenon occurs. Deep propagation can take the form of multipaths, which come together at convergence zones, or the propagation may simply be confined to the deep sound channel.

extremely involved and can be regarded as an independent subject of study.

V. APPLICATIONS

A. Active Sonar

Two important applications of the principles of acoustic signal processing are the active sonar problem, which is an example of a forward problem, and large-scale ocean tomography, which is an example of an inverse problem. (This example is based on the work of Sato and Shiraki.) For the active sonar problem, a simple example using a sine wave pulse against a moving target, which can be assumed to be made up of scattering "points," will be investigated.

In order to use a replica correlator, the stored replica must be assumed to be identical to the transmitted signal. This assumption is useful since, even though the target is not a true point, the actual scattering produces an echo that can be approximated by scattering from a set of points. Thus, the echo can be assumed to be a superposition of replicas of the transmitted signal. The total receiver consists of (1) an acoustic array, (2) a beam former, (3) a quadrature detector, and (4) a threshold operation. This is shown in Fig. 16, where the beam former is identical to that shown in Fig. 8 and the quadrature receiver is that shown in Fig. 4. If there are N beam outputs and M spectral outputs from the spectral analyzers (fast Fourier transforms), then there are $N \times M$ outputs of the beam former. For each of these outputs, there must be a quadrature receiver. This process can be performed digitally. The boxes labeled "A/D converter" sample the output voltage of each transducer at a rate equal to at least twice the bandwidth of the received signal. This is referred to as the Nyquist rate and is the minimum sample rate necessary to extract all of the information from the signal.

The performance of the system can be conveniently discussed by means of the ambiguity function, which is

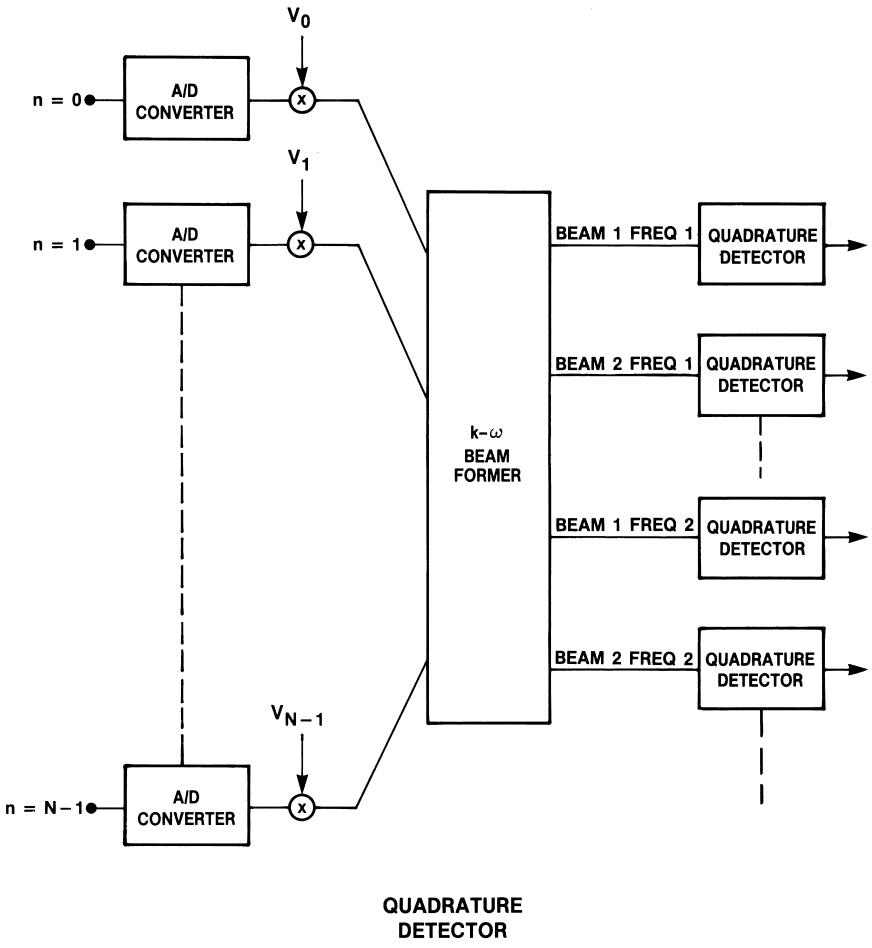


FIGURE 16 Processor for active sonar receiver. The $k-\omega$ beam former is that shown in Fig. 8, and the quadrature detector is that shown in Fig. 4.

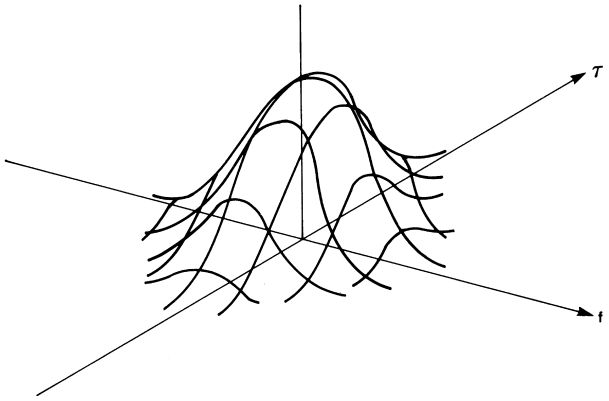


FIGURE 17 Ambiguity function for single-frequency pulse.

the output of the quadrature receiver as a function of the assumed delay time and the assumed frequency. This is illustrated in Fig. 17. The width of the ambiguity function along the τ axis is a measure of the capacity of the system to resolve the range of the target and is approximately equal to T, the duration of the pulse for a sine wave pulse. On the other hand, the width of the function along the frequency axis is approximately equal to the reciprocal of the pulse duration. Thus, for this example, one cannot achieve arbitrarily good range and Doppler resolution with a single pulse. The pulse duration is therefore an important design parameter.

Another important design parameter is the operating frequency. This also imposes conflicting requirements. For example, the capacity of the system to resolve bearing angle is determined partly by the beamwidth of the receiving array response function. This beamwidth is approximated by the reciprocal of the aperture length in wavlengths, as discussed in Section III. Usually the aperture size of the array is constrained by the size of the platform. This maximum aperture size thus determines the lowest frequency that can be used. Of course, one could simply increase the frequency, thus achieving any desired narrowness of the receiving array beam, but as discussed in Section IV, the higher the frequency, the greater is the absorption of the acoustic energy by the seawater.

In the above example, no consideration was given to reverberation. If this were a problem, the single-frequency pulse could be replaced by an FM slide, with the appropriate changes made in the matched filter.

B. Oceanic Tomography

An example of acoustic signal processing as applied to an inverse problem is that of oceanic tomography. Large-scale experiments have been performed that have used the time delay incurred in transmissions between pairs of transducers. This method recommends itself, since the number of data is equal to the product of the number of transmitters and the number of receivers. In practice, the problem is quite difficult, since as seen in Section IV, the long-range propagation can be very complex due in part to the existence of multiple paths (sometimes called multipaths). The time delays along these multiple paths are then considered as individual data and reflect in part the speed-of-sound profile. The method has been used to determine the large-scale speed-of-sound map of the ocean. This in turn yields information on temperature-driven gyres. Thus, such an inverse problem that directly measures time delay gives indirect information on temperature distribution and thus gyre structure. The effect of the currents themselves does not strongly influence the time delay measurements, since the temperature is the main determinant of the sound speed, at least above the deep sound channel.

In order to reduce the complexity of the example it will be assumed that the temperature is a constant and that the process is intended to measure the current speed itself via time delay measurement. Although a much simpler problem than that discussed above, it clearly demonstrates the inverse nature of the process.

Consider the region being subjected to measurement to be surrounded by N transducers, which are both transmitters and receivers. Thus, there are $(N^2 - N)/2$ paths. In practice, all of these paths may not be useful due to constraints on the positioning of the transducers, but here we will assume that they are. Let the region in which the flow is to be determined be divided into subregions. Labeling the path by i and the subregion by j, we give the sound speed along path i in region j by

$$c_{ij} = c_0 \pm (U_{xj} \cos \delta_i + U_{yj} \sin \delta_i), \tag{68}$$

where U_{xj} is the x component of U_j, U_{yj} the y component of U_j, U_j the current speed in subregion j (assumed constant over the subregion), δ_i the angle of the path from the x axis, and c_0 the speed of sound (assumed to be the same in all subregions). The travel time of the acoustic pulse along the ith path is then given by summing over the regions transversed by path i:

$$\tau_{i\pm} = \sum_j \frac{L_{ij}}{C_0 \pm (U_{xj} \cos \delta_i + U_{yj} \sin \delta_i)}. \tag{69}$$

Here, L_{ij} is the length of path i in subregion j; the plus and minus subscripts indicate sound propagation with or against the flow. Assuming c_0 to be much greater than the current speed, Eq. (69) can be approximated by

$$\tau_{i\pm} \simeq \sum_j \frac{L_{ij}}{C_0^2} [C_0 \mp (U_{xj} \cos \delta_i + U_{yj} \sin \delta_i)]. \tag{70}$$

The difference between the backward and forward travel times is defined as

$$T_i = \tau_{i-} - \tau_{i+} = \sum_j \frac{2L_{ij}}{C_0^2}[U_{xj}\cos\delta_i + U_{yj}\sin\delta_i]. \tag{71}$$

If there are Q subregions, there are $2Q$ current velocity components to be determined. Hence, $2Q$ acoustical paths are required. Clearly, Q cannot be greater than $(N^2 - N)/4$. Since the geometry of the situation is taken to be known, the values of the δ_i are known along with the L_{ij}. Defining

$$A_{ij} = \frac{2L_{ij}}{C_0^2}\cos\delta_i, \tag{72}$$

$$B_{ij} = \frac{2L_{ij}}{C_0^2}\sin\delta_i, \tag{73}$$

we can rewrite Eq. (71):

$$T_i = \sum_j [A_{ij}U_{xj} + B_{ij}U_{yj}]. \tag{74}$$

The index i ranges from 1 to $2Q$, and the index j ranges from 1 to Q. Equation (74) can be rewritten in matrix form as

$$T_i = K_{ij}U_j. \tag{75}$$

From Eqs. (72) to (75) we see that

$$K_{ij} = (A_{ij}, B_{ij}), \tag{76}$$

$$U_j = (U_{xj}, U_{yj}). \tag{77}$$

Equation (75) has the same form as Eq. (1) in Section I. And as is usually the case in inverse problems, the matrix $\mathbf{K}$ must be inverted to yield the solution $\mathbf{U}$. That is, the solution vector of flow velocities is given by

$$\mathbf{U} = \mathbf{K}^{-1}\mathbf{T}. \tag{78}$$

The process by which one inverts Eq. (74) [or Eq. (75)] could be any one of several schemes, the selection of which can depend on the physical situation itself, the noise content of the data, the computational load required by the inversion process, and whether the matrix is well conditioned. Whatever technique is chosen, the solution will allow a two-dimensional map of the flow field of the region of measurement. The accuracy of the solution will depend on the performance of the algorithm chosen for the solution, how well posed the problem is, the noise content of the data, and the fineness of the mesh (i.e., how many subregions the area of interest was broken into).

An example of how this problem could become ill posed is given by the case where the currents are very small. In this case, since the measurement vector $\mathbf{T}$ has elements that are differences between measured values of travel time and these travel times will differ by very little, large errors could occur in the solution vector $\mathbf{U}$, given sufficient noise or insufficient measurement accuracy. Also, depending on the values of δ_i and L_{ij} in Eqs. (72) and (73), the matrix may or may not emphasize or "amplify" these errors. Thus, great care must be taken in designing the algorithm given the accuracy of the measurements. And it may, in fact, be the case that no algorithm will suffice if the measurements are not accurate enough.

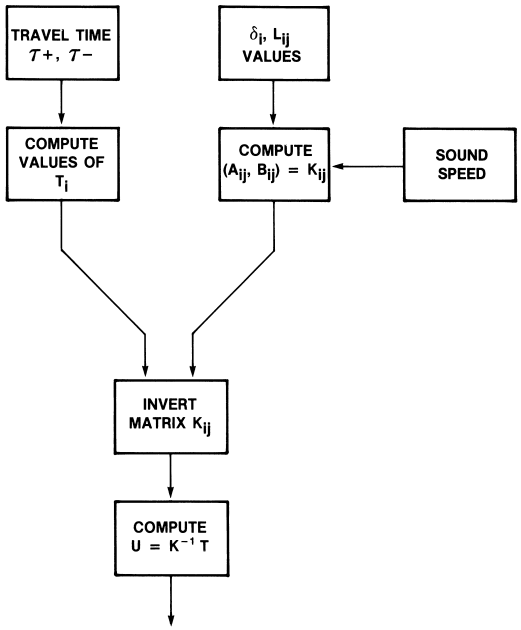


FIGURE 18 Flow chart for inverse-problem algorithm.

The implementation of such a scheme, as is the case with just about every other inverse problem, is software intensive. A flow chart of how Eq. (78) would be solved for the vector $\mathbf{U}$ is shown in Fig. 18.

Finally, it should be emphasized that every inverse problem necessarily contains a forward problem. That is, the inputs to an inverse problem contain measured data. These data are determined by a detection and estimation process. In the foregoing problem, the travel time data would most likely be determined by transmitting pulses whose spectral and temporal characteristics would be specified by the forward problem of detecting the pulses in the given noise environment and estimating the time delays between their transmission and reception.

VI. RECENT DEVELOPMENTS

A. Model-Based Passive Localization

An approach to signal processing that has been receiving increased attention recently is model-based signal processing. It is basically a means of including *a priori*

information in the processing scheme. An example of model-based signal processing is the passive localization scheme sometimes referred to as "matched-field" processing. It is basically a problem in the estimation of the range, depth, and bearing of an acoustic point source in a parallel-plane waveguide by using a propagation model in conjunction with measurements from a receiving array. As such, it is another example of an inverse problem. The name arises from the fact that the source coordinates used in the propagation model are varied until the acoustic field predicted at the array by the model yields a best "match" with the measured array data. The source coordinates producing this best match are then taken as the estimated source coordinates. The propagation model provides the mechanism for including the *a priori* environmental information in the processing scheme. In the case of underwater acoustics, the relevant environmental information includes the sound velocity profile, the ocean depth, and the ocean bottom properties. In this discussion, in the interest of simplicity, the problem shall be limited to the two-dimensional case of range and depth only. The propagation model is the so-called normal-mode model, which is the solution to the acoustic wave equation for the field radiated by a single-frequency point source in a parallel-plane waveguide made up of the ocean surface and bottom. It is valid for shallow water situations, i.e., those cases where the depth in units of the acoustic wavelength is not excessively large. The "modal functions" are the parts of the solution that satisfy the boundary conditions at the surface and bottom of the ocean. There is a finite number of these modal functions, expressed as $Z(z)$, each of which propagates at its own unique phase velocity. The total field is the weighted sum of these modal functions and is given by

$$P(r, z) = \sum_{m=1}^{M} Z_m(z) Z_m(z_0) \frac{e^{ik_m r}}{\sqrt{k_m r}}. \tag{79}$$

In Eq. (79) r is the horizontal range from the point source, assumed to be at $r = 0$; k_m is the so-called horizontal wave number associated with the mth mode; and z_0 is the depth of the source measured positively downward from the surface. $P(r, z)$, then, is the field at depth z from the surface and distance r from an acoustic point source which is at depth z_0.

As an example, consider the situation shown in Fig. 19, where it is seen that a vertical array is used to obtain the measurements of the acoustic field. Other array configurations could have been used. Assuming the number of hydrophones N to be greater than M, the number of modes, the problem is overdetermined, and an estimate of the modal amplitudes, which upon inspection of Eq. (79) are given by $Z_m(z_0)\, e^{ik_m r}/\sqrt{k_m r}$, can be found by the method of least squares. Rewriting Eq. (79) in matrix formalism, we have

$$P_n = Z_{n,m} \beta_m, \tag{80}$$

where P_n is the nth element of the vector of field measurements at the N hydrophones of the vertical array, $Z_{n,m}$ is the (n, m)th element of the modal matrix, and β_m is the mth element of the vector of M modal amplitudes. Thus, the vector $\mathbf{P}$ represents the measurement, the matrix $\mathbf{Z}$ represents the model, and the vector β carries the information concerning the range and depth of the source. The solution of the problem proceeds as follows. Given the sound velocity profile, the ocean depth, and the material properties of the ocean bottom, the wave equation can be solved for the modal functions $Z_m(z)$ and the horizontal wave numbers k_m. The least-squares estimate for β is found by determining the vector $\hat{\beta}$ that satisfies

$$\sum_{n=1}^{N} (P_n - Z_{n,m} \hat{\beta}_m)^2 = \min. \tag{81}$$

Here, P_n represents the nth component of the vector of hydrophone measurements. It can be shown that, for the

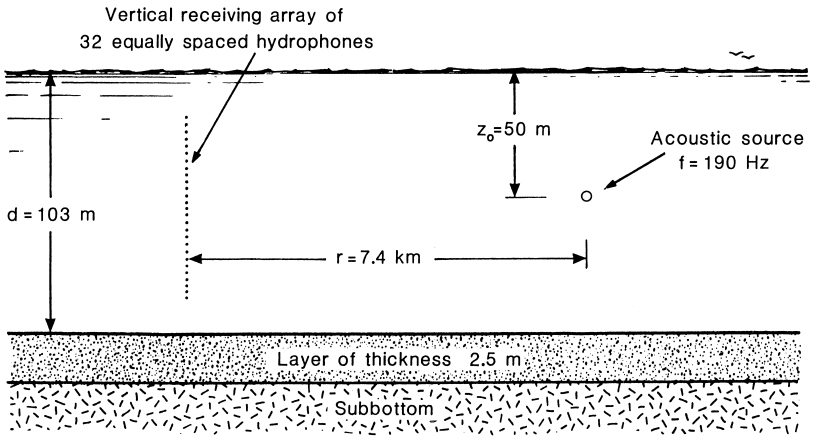


FIGURE 19 Source-receiver configuration for model-based passive localization experiment.

case of $N > M$, the solution of Eq. (81) for $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{P}. \tag{82}$$

The quantity $(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ is called the pseudoinverse of the matrix $\mathbf{Z}$ and $\mathbf{Z}^T$ is its transpose. When $N = M$, as expected, $(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ reduces to $\mathbf{Z}^{-1}$. The range and depth estimates are now found by numerically solving the least-squares problem given by

$$\sum_{m=1}^{M}\left(\hat{\beta}_m - Z_m(z_0)\frac{e^{ik_m r}}{\sqrt{k_m r}}\right)^2 = \min \tag{83}$$

over a grid of range and depth values. As can be seen, this requires knowledge of the horizontal wave numbers. That range–depth pair that produces the global minimum of Eq. (83) is taken as the solution. Of course, the problem could have been solved by directly numerically solving Eq. (81). The advantage of the method used here is that the problem is solved in mode space instead of hydrophone space. This allows any particular modes that might be disturbed by mismatches between the model and the actual physical propagation channel to be removed.

The results of applying this technique to data from an experiment based on the scenario depicted in Fig. 19 are shown in Fig. 20. This solution was obtained only after removing the two highest modes, which were modes number 8 and 9, since the propagation channel supported nine modes. This points out the utility of the mode filtering property of the solution method outlined above. In Fig. 20, instead of directly plotting the right-hand side of Eq. (83), the surface has been inverted to provide a surface with a maximum at the minimum of Eq. (83). In the case of white Gaussian noise, the solution of Eq. (83) would be the maximum-likelihood solution so that the maximum of the surface in Fig. 20 would then correspond to this maximum-likelihood solution. The experimental data were taken on a vertical array with 32 receiving elements with 2 m spacing. The distance to the source was 7.4 km. The source frequency was 190 Hz, and the source depth was 50 m. The region where the data were taken is an area of the Mediterranean sea north of the island of Elba where the depth is 103 m and the bottom is quite flat. In Fig. 20 it can be seen that the range estimate is quite accurate, but there is a bias in the depth estimate. This is most likely due to the fact that the propagation model does not provide a perfectly faithful representation of the actual physical situation, as evidenced by the necessity of removing two modes in order to obtain a solution. In matched-field processing, this is sometimes referred to

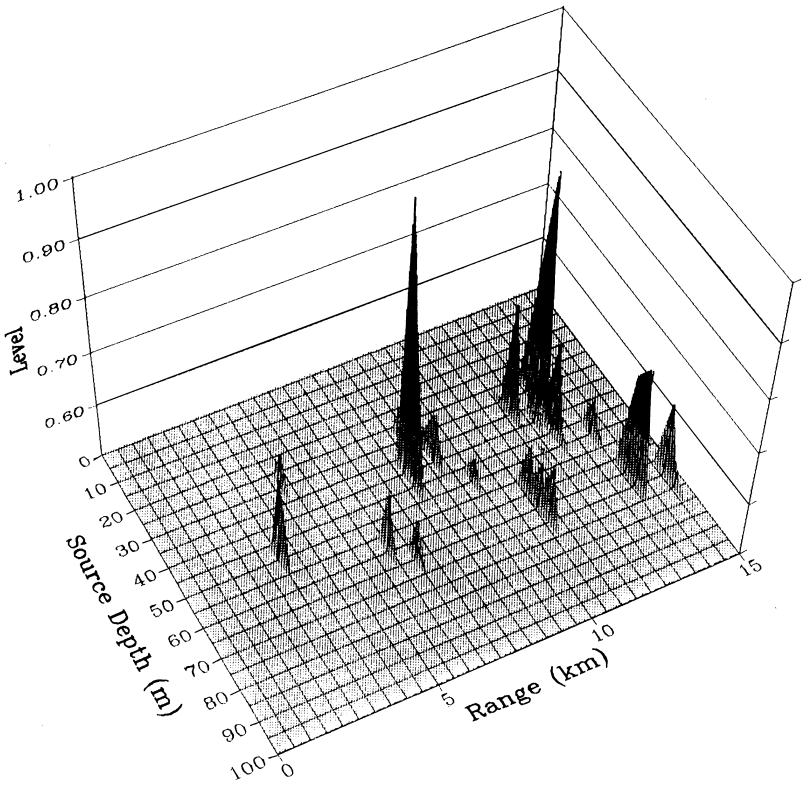


FIGURE 20 Likelihood surface using data from the model-based passive localization experiment. The true coordinates are: Range = 7.4 km and depth = 50 m.

as the "mismatch" problem. For our purposes, however, it provides another good example of an ill-posed inverse problem whereby small errors in the input parameters can cause appreciable errors in the results.

An approach to the mismatch problem is to consider the problematical model parameters as parameters to be estimated along with the source coordinates in a so-called multivariate estimation problem.

There are many other applications of model-based processing. One example is bearing estimation with a towed array where a specific model of the signal is introduced. When the forward motion of the array is included in the model, the performance is improved through the passive synthetic aperture effect. That is, the array appears to be much longer than it actually is.

B. Higher-Order Spectral Analysis

At a more sophisticated level, there are more subtle characteristics of an acoustic signal that can be exploited for signal processing purposes. Two particular types of stochastic signals that carry useful information are the cases of transient signals and nonlinear signals. An example of a nonlinear signal would be a random signal that passes through a nonlinear transform. This could be a mechanical noise or vibration driving a ship's hull plate, where the hull plate obeys a nonlinear, say quadratic, Hooke's law. The acoustic radiation from this plate would then contain intermodulation frequencies with stable phases generated from the frequency components of the Gaussian signal by the nonlinearity. One method of dealing with such signals is by higher-order spectral analysis.

The simplest example of higher-order spectral analysis is the bispectrum. Just as the power spectral density of a random signal is the Fourier transform of the autocorrelation function, the bispectrum is the double Fourier transform of the third-order correlation function. That is, the power spectral density is given by

$$P(\omega) = \int_{-\infty}^{+\infty} d\tau e^{-i\omega\tau} R(\tau), \tag{84}$$

where

$$R(\tau) = E\{x(t)x^*(t+\tau)\}, \tag{85}$$

and the bispectrum is given by

$$B(\omega_1, \omega_2) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} d\tau_1\, d\tau_2\, e^{-i(\omega_1\tau_1 - \omega_2\tau_2)} R(\tau_1, \tau_2), \tag{86}$$

where

$$R(\tau_1, \tau_2) = E\{x(t)x(t+\tau_1)x^*(t+\tau_2)\}. \tag{87}$$

The symbol E denotes the expected value and $x(t)$ is the signal. The asterisk denotes the complex conjugate. In this development, the signal is assumed to be stationary; therefore $R(\tau)$ and $R(\tau_1, \tau_2)$ do not depend on t. A major difference between the bispectrum and the power spectral density is that the bispectrum contains phase information from the signal where the ordinary power spectral density does not. It is this property that allows the bispectrum to indicate the existence of the intermodulation terms mentioned earlier in regard to nonlinear signals.

The quantity that is computed for signal processing purposes is the "skewness," which is given by

$$\Gamma(\omega_1, \omega_2) = \frac{|B(\omega_1, \omega_2)|}{[P(\omega_1)P(\omega_2)P(\omega_1+\omega_2)]^{1/2}}. \tag{88}$$

The quantity $P(\omega)$ is the power spectral density as defined in Eq. (84). If the signal is Gaussian, the expected value of $\Gamma(\omega_1, \omega_2)$ is zero. This is because the Gaussian probability density function is symmetric, and the skewness of any symmetric probability density function is zero. If the signal is non-Gaussian but linear, the expected value of $\Gamma(\omega_1, \omega_2)$ is nonzero but independent of frequency. That is, it is a constant on the ω_1, ω_2 plane. However, if the signal is nonlinear, the expected value of $\Gamma(\omega_1, \omega_2)$ depends on ω. That is, it is *not* a constant on the ω_1, ω_2 plane. This is a result of the existence of the intermodulation terms mentioned previously.

The derivation of Eq. (88) is complicated and will not be discussed here, but as mentioned previously, it assumes the signal to be stationary. This assumption is the reason that the bispectrum depends only on two frequencies instead of three. In the case of nonstationarity, the bispectrum would depend on three frequencies since the dependence of the correlation function on t would necessitate a triple Fourier transform, thus introducing a third frequency variable. This has the effect of enlarging the principle region of the skewness function on the ω_1, ω_2 plane. Because of this, the existence of transients (or nonstationarity) can be tested for by constructing an estimate of the skewness function in this supplemental region of the ω_1, ω_2 plane.

Thus, by constructing an estimate of the skewness function from the bispectrum, a stochastic time series can be tested for non-Gaussianity, nonlinearity, and stationarity. The stationarity test, of course, allows the signal to be tested for the presence of transients.

The bispectrum has application in many areas of acoustic signal processing. In underwater acoustics, it can be used to detect submarines, since their propulsion machinery, propellers, and associated hydrodynamic phenomena produce nonlinear acoustic radiation, whereas the ambient acoustic radiation can be expected to be Gaussian and therefore linear. In the area of nondestructive testing,

defects in machinery such as gear damage can be detected by bispectral analysis of the noise radiated by the machinery. This is a consequence of the fact that the statistics of the gear noise become either more or less Gaussian in comparison to the noise radiated by "good" gears. Also, since the echo of an active sonar ping can be considered to be a transient signal, the bispectrum can be used as an active sonar detector which, unlike the matched filter, does not require *a priori* knowledge of the configuration of the echo.

Although only the bispectrum has been discussed here, it is not the only example of higher-order spectral analysis. Work is currently being carried out on the trispectrum, which is defined as the *triple* Fourier transform of the *fourth*-order correlation function. As might be expected, the trispectrum has the ability to extract even more information from the acoustic signal because it exploits higher order moments of the signal than the bispectrum does.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • ACOUSTICS, LINEAR • PHYSICAL OCEANOGRAPHY • RADAR • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL • ULTRASONICS AND ACOUSTICS • UNDERWATER ACOUSTICS

BIBLIOGRAPHY

Behringer, D., Birdsall, T., Brown, M., Cornuelle, B., Heinmiller, R., Knox, R., Metzger, K., Munk, W., Spiesberger, J., Spindel, R., Webb, D., Worcester, P., and Wunsch, C. (1982). *Nature* (*London*) **299,** 121.

Burdic, W. S. (1984). "Underwater Acoustic System Analysis," Prentice-Hall, Englewood Cliffs, NJ.

Candy, J. V. (1968). "Signal Processing: The Model-Based Approach," McGraw-Hill, Singapore.

Candy, J. V., and Sullivan, E. J. (1989). *J. Acoust. Soc. Am.* **85**(6), 2472.

Cox, H. (1973). *J. Acoust. Soc. Am.* **54,** 1289.

Hinich, M. J., Marandino, D., and Sullivan, E. J. (1989). *J. Acoust. Soc. Am.* **85**(4), 1512.

Hinich, M. J., and Sullivan, E. J. (1989). *J. Acoust. Soc. Am.* **85**(1), 214.

Hudson, J. E. (1981). "Adaptive Array Principles," Peter Peregrinus, London.

Metherell, A. F., ed. (1980). "Acoustical Imaging," Vol. 8, Plenum, New York.

Sarkar, T. K., Weiner, D. D., and Jain, V. K. (1981). *IEEE Trans. Antennas Propag.* **AP-29**(2), 373.

Sato, T., and Shiraki, M. (1984). *J. Acoust. Soc. Am.* **76**(5), 1427.

Sullivan, E. J., and Middleton, David (1993). *IEEE J. Oceanic Eng.* **18**(3), 156.

Sullivan, E. J., and Candy, J. V. (1997). *J. Acoust. Soc. Am.* **102**(5, Part 1), 2809.

Tolstoy, A. (1993). "Matched Field Processing For Ocean Acoustics," World Scientific, River Edge, NJ.

Urkowitz, H. (1983). "Signal Theory and Random Processes," Artech House, Dedham, MA.

Whalen, A. D. (1971). "Detection of Signals in Noise," Academic Press, New York.



Signal Processing, Analog

W. K. Jenkins
University of Illinois

I. Introduction
II. Background Mathematics
III. Analog Filters Required in A/D and D/A Conversion
IV. Analog Computing
V. Characteristics of Analog VLSI Circuits
VI. Switched-Capacitor Circuits
VII. Switched Analog Filters
VIII. Optical Signal Processing

GLOSSARY

Active filter Class of frequency-selective analog filters constructed from resistors, capacitors, and operational amplifiers.

Analog computer Programmable analog circuit simulator that consists of high-quality operational amplifiers, high-precision resistors and capacitors, and analog multipliers that can be conveniently patched together to produce real-time solutions for both linear and nonlinear differential equations.

Chirp waveform Sinusoidal pulse with a quadratic phase (linear frequency) modulation that is often used in pulsed-Doppler radars.

Compression filter Matched filter for a prescribed waveform such as a chirp, often used as detection filters in radar receivers.

Homogeneous differential equation Unforced differential equation that results from setting the forcing function to zero in a linear constant-coefficient differential equation.

Ideal reconstruction formula Infinite summation of weighted $(\sin x)/x$ basis functions capable of perfectly reconstructing an analog signal from a set of samples taken at or above the Nyquist sampling rate.

Nonhomogeneous differential equation Linear constant-coefficient differential equation with the forcing function activated, that is, not set to zero.

Operational amplifier Differential voltage amplifier characterized by a very high linear voltage gain, a very high input impedance, and a very low output impedance.

Passive filter Class of frequency-selective analog filters constructed from resistors, capacitors, and inductors.

Postfilter (reconstruction) Analog filter cascaded with a digital-to-analog (D/A) converter to smooth the staircase reconstruction error and compensate for the

$(\sin x)/x$ distortion inherent in the D/A reconstruction process.

Prefilter (antialiasing) Analog low-pass filter used before an analog-to-digital (A/D) converter to remove all frequency content above the half-sampling frequency and prevent aliasing distortion.

Synthetic aperture radar Microwave radar carried on an airplane or orbiting satellite which collects radar returns from many different spatial locations along the flight path and produces a high-resolution terrain image by combining many returns through signal processing.

Switched-capacitor filter Class of frequency-selective analog filters constructed from capacitors, transistor switches, and operational amplifiers (op amps) and that are particularly well suited for monolithic fabrication.

Zero-input response Response due to initial energy storage in the system of a linear system with a zero input applied.

Zero-state response Response of a linear system to an input applied when the system is in the zero state (no initial energy storage).

THE TERM *analog signal processing* has been used to distinguish between traditional continuous-time signal-processing techniques and techniques in digital signal processing. In this article, the term *analog* is used synonymously with *continuous time*; a *continuous-time* (CT) signal is a function $s(t)$ that is defined for all time t contained in some interval on the real line. In this context the independent variable is usually called time, because in many signal-processing applications the functional relationships of interest are indeed dependent on time as the independent variable. However, in other applications the independent variable may represent different physical quantities such as spatial displacement or environmental temperature. The term analog signal processing refers to the representation and processing of all signals that satisfy the usual continuous-time definition, regardless of the specific physical interpretation of the independent variable.

I. INTRODUCTION

Analog signals are processed by specially designed devices, circuits, or systems to extract parametric information or to alter the characteristics of the input signal in some prescribed way. Spectral shaping is probably the most common form of analog signal processing and is typically done with either passive (containing resistors, capacitors, and inductors) or active [containing resistors, capacitors, and operational amplifiers (op amps)] analog filters. The active filter is probably the most common analog signal processor in use today. During the past few years, switched-capacitor filters have become popular as high-precision replacements for certain types of active filters that are to be realized in monolithic form as part of a more extensive integrated system. Switched-capacitor circuits, which contain capacitors, transistor switches, and operational amplifiers, have been successfully used for analog pre- and postfilters required at the interface between analog and digital systems. This subject is treated in more detail in Section VI. Both active filters and switched-capacitor filters are closely related historically to the analog computer, from which the term analog first originated.

Analog signals can be processed by many different types of devices, some of which are surface acoustic wave (SAW) filters, charge-coupled devices (CCDs), high-frequency distributed-parameter circuits, or optical circuits, to name only a few. Since it will not be possible to discuss all the different classes of analog processing techniques and devices, and because many of these are treated elsewhere, attention will be devoted to reviewing important mathematical tools used in analog signal analysis, discussing the connection between modern analog signal processing and its origins in the analog computer, and examining several particularly important historical topics in analog signal processing. These special topics include switched-capacitor filters, electronically tunable analog filters, nonlinear modeling by means of the analog computer, and image reconstruction for synthetic aperture radar by optical processing techniques.

II. BACKGROUND MATHEMATICS

Linear analog signal processors are characterized by the familiar mathematics of differential equations, continuous convolution operators, Laplace transforms, and Fourier transforms. Consider the single-input, single-output system shown in Fig. 1, where $y(t) = \mathbf{S}\{x(t)\}$, $x(t)$ is the analog input signal, and $y(t)$ the analog output.

The following definitions are commonly used to define the properties of linearity, time-invariance, causality, and stability:

1. *Linearity*. $\mathbf{S}\{\cdot\}$ is linear if and only if $\mathbf{S}\{ax_1(t) + bx_2(t)\} = a\mathbf{S}\{x_1(t)\} + b\mathbf{S}\{x_2(t)\}$, where a and b are scalar constants and $x_1(t)$ and $x_2(t)$ are two arbitrary input functions.

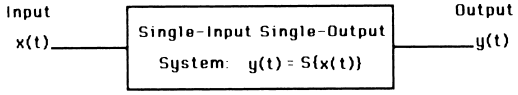


FIGURE 1 Single-input, single-output system $\mathbf{S}\{\cdot\}$.

2. *Time-invariance*. $\mathbf{S}\{\cdot\}$ is time-invariant if and only if $y(t-t_0)=\mathbf{S}\{x(t-t_0)\}$ for all real values of t_0.

3. *Causality*. $\mathbf{S}\{\cdot\}$ is causal if and only if $y(t)$ depends only on values of $x(s)$ at times $s \le t$.

4. *Stability (BIBO)*. $\mathbf{S}\{\cdot\}$ is stable in the bounded input–bounded output sense if and only if the output $y(t)$ remains bounded for all inputs $x(t)$ that are bounded.

5. *Stability (asymptotic)*. $\mathbf{S}\{\cdot\}$ is stable in the asymptotic sense if and only if whenever $x(t)=0$ for $t \ge t_0$, $y(t)$ approaches zero as t approaches infinity.

Whenever $\mathbf{S}$ is linear and time-invariant, it is possible to express the zero-state response $y(t)$, that is, the response when the system has no initial energy storage, due to an arbitrary input $x(t)$ in terms of the convolution of $h(t)$ with $x(t)$ as expressed by Eq. (1), where $h(t)=\mathbf{S}\{\delta(t)\}$ is an analog impulse function (Dirac delta function) and $h(t)$ is the impulse response of $\mathbf{S}$.

$$y(t) = h(t) * x(t) = \int_{-\infty}^{+\infty} h(t)x(t-\tau)\,d\tau. \qquad (1)$$

Note that Eq. (1) characterizes only the zero-state component of the response, that is, only that part which is excited by the application of the input. In general, the zero-input response (i.e., that part due to initial energy storage in the system) must be accounted for in a different manner.

If $\mathbf{S}$ is causal, $h(t)=0$ for $t<0$, and the lower limit in Eq. (1) becomes zero. Also, if $x(t)=0$ for $t<0$, then the upper limit on the convolution becomes t. The convolutional property of linear time-invariant systems is an important fundamental property that can be summarized by the following theorem.

Fundamental theorem of linear systems. The zero-state output of a linear time-invariant analog system in response to an arbitrary input signal $x(t)$ can be expressed as the convolution of $x(t)$ with the impulse response $h(t)$ of the system, that is, $y(t)=h(t)*x(t)$.

The fundamental theorem of linear systems guarantees that as long as the impulse response of the system is known and as long as the system is linear and time-invariant, then the output of the system can be computed for any input signal without having any detailed knowledge about the internal structure of the system.

A. Characterization by Differential Equations

There are important subclasses (such as passive and active analog filters) of linear time-invariant systems that can be characterized by an Nth-order differential equation expressed in terms of the input $x(t)$ and the output $y(t)$, as illustrated by

$$d^N y(t)/dt^N + \cdots + a_1 dy(t)/dt + a_0 y(t) = b_M d^M x(t)/dt^M + \cdots + b_1 dx(t)/dt + b_0 x(t). \qquad (2)$$

Time-invariance guarantees that the coefficients in Eq. (2) are constants. In certain situations it may be difficult to describe a system with a single differential equation, but a set of lower order coupled differential equations may be more convenient. The traditional normal-form state equations of the form $d\mathbf{X}(t)/dt = \mathbf{A}\mathbf{X}(t) + \mathbf{B}\mathbf{U}(t)$ are simply a standard set of coupled first-order differential equations used to describe an arbitrary linear time-invariant system in a standard language. Once a system has been characterized in a standard way, then standard mathematical techniques can be used to find solutions for time response, frequency response, transfer functions, stability, etc.

Fortunately, the vast majority of analog filters can be described with differential equation approaches, and hence they can be analyzed rigorously with the above approach. This is true because linear capacitors and inductors satisfy strict first-order differential laws of the form $y(t)=\boldsymbol{K}\,dx(t)/dt$ between appropriate sets of terminal variables. The inductor law is a result of Faraday's natural law of induced voltage, while the capacitor law is due to the fact that current is the rate of flow of charge in a device. Since resistors, op amps, and ideal transformers all satisfy algebraic relationships among their terminal variables, it follows that all ideal passive and active analog filters can be characterized by appropriate sets of differential equations, because they are represented by collections of these elementary properties whose natural form is preserved by Kirchoff's laws of interconnection.

Although the entire subject of differential equations cannot be adequately reviewed here, there are several issues that should be mentioned. Linear constant-coefficient differential equations can be solved either by classical techniques or by more modern transform techniques. The classical approach separates the problem into homogeneous and nonhomogeneous equations, finds solutions to these individually, and then combines them to produce the complete solution. The classical approach is intuitive and convenient in that it breaks one large problem into several smaller ones, but it also involves a certain degree of "guesswork" that the neophyte always finds disturbing. Also, the boundary values, or initial conditions, must be patched into the general solution at the very end to finally get a specific solution. In contrast, the transform method is a powerful approach that includes initial conditions at the very beginning of the procedure and produces a final result through a sequence of mechanical steps that do not require much insight or intuition. In other words, a solution can often be obtained by sheer persistence and hard work. The disadvantage of the transform approach is that, although

the mathematical manipulations may be straightforward in principle, the actual computations often become extremely tedious, and it is not always easy to justify the result intuitively without further work. The experienced analyst finds it useful to keep both the classical and transform methods nearby and to use them interchangeably as necessary. It is simply misleading to claim that any one method of solution will result in the best approach when applied over a broad class of problems.

There is also a common misconception that the complementary solution (obtained from the homogeneous equation) and the particular solution (obtained from the nonhomogeneous equation) in a classical solution are the same as the zero-input response and the zero-state response from a modern state-space solution. That this is not true can be seen in Eq. (1), where it is apparent that the application of an input signal to a system that was initially in zero state may very well lead to the excitation of transient natural response terms. In the state-space approach, these transients arise as part of the zero-state response, since they do not depend in any way on initial energy storage in the system. However, in the classical solution, these terms turn up as part of the complementary solution, since the complementary solution contains all natural response terms, regardless of whether they are excited by the input or are due to initial energy storage. Another way to see this is to note that the complementary solution does indeed depend on the particular solution, because the boundary values are used in the total solution after the various components have been added together.

B. The Role of the Laplace Transform in Analog System Analysis

In the preceding section there was a discussion of transform methods for solving differential equations, without explicitly mentioning which transform could be used. The most familiar transform used in analog signal analysis is the Laplace transform, as defined by

$$F(s) = \mathcal{L}\{x(t)\} = \int_0^{+\infty} f(t)e^{st}\,dt \tag{3a}$$

$$x(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi}\int_{\sigma-j\infty}^{\sigma+j\infty} F(s)e^{-st}\,ds. \tag{3b}$$

The region of convergence of $F(s)$ is the set $R = \{s|F(s) \neq \infty\}$ (i.e., it is the set of all s such that Eq. (3a) converges). In Eq. (3b), the contour of integration must be selected to lie entirely within R. Since the lower limit on the integral in Eq. (3a) is zero, this form is known as the one-sided (unilateral) Laplace transform to distinguish it from the two-sided (bilateral) form. In analog system analysis, the Laplace transform is typically used to convert differential equations into algebraic equations, which can then be solved by algebraic techniques. The time-domain solution can then be found by computing the inverse Laplace transform by means of Eq. (3b) or by the use of a Laplace transform table. A list of Laplace transforms is given in Table I for some common functions used frequently in analog signal processing.

Linearity is an important property of the Laplace transform, a property that it inherits from the linear property

TABLE I Laplace Transforms of Some Common Functions[a]

$f(t)$[b]	$F(s)$
$u(t)$	$\frac{1}{s}$
t	$\frac{1}{s^2}$
$\frac{t^{n-1}}{(n-1)!}$ $n =$ integer	$\frac{1}{s^n}$
e^{at}	$\frac{1}{s-a}$
te^{at}	$\frac{1}{(s-a)^2}$
$\frac{1}{(n-1)!}t^{n-1}e^{at}$	$\frac{1}{(s-a)^n}$
$\frac{1}{a-b}(e^{at}-e^{bt})$	$\frac{1}{(s-a)(s-b)}$
$\frac{e^{-at}}{(b-a)(c-a)} + \frac{e^{-bt}}{(a-b)(c-b)} + \frac{e^{-ct}}{(a-c)(b-c)}$	$\frac{1}{(s+a)(s+b)(s+c)}$
$1-e^{+at}$	$\frac{-a}{s(s-a)}$
$\frac{1}{\omega}\sin\omega t$	$\frac{1}{s^2+\omega^2}$
$\cos\omega t$	$\frac{s}{s^2-\omega^2}$
$1-\cos\omega t$	$\frac{\omega^2}{s(s^2-\omega^2)}$
$\sin(\omega t+\theta)$	$\frac{s\sin\theta-\omega\cos\theta}{s^2-\omega^2}$
$\cos(\omega t+\theta)$	$\frac{s\cos\theta-\omega\sin\theta}{s^2-\omega^2}$
$e^{-\alpha t}\sin\omega t$	$\frac{\omega}{(s+\alpha)^2+\omega^2}$
$e^{-\alpha t}\cos\omega t$	$\frac{s-\alpha}{(s+\alpha)^2-\omega^2}$
$\sinh\alpha t$	$\frac{\alpha}{s^2-\alpha^2}$
$\cosh\alpha t$	$\frac{s}{s^2-\alpha^2}$

[a] From Van Valkenburg, M. E. (1974). "Network Analysis," 3rd ed., p. 192. © 1974. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

[b] All $f(t)$ should be thought of as being multiplied by $u(t)$ (i.e., $f(t) = 0$ for $t < 0$).

of the integral operator used in its definition. Two other important properties are the following:

Integral theorem of the Laplace transform. If $f(t)$ is a Laplace transformable function with transform $F(s)$, then

$$\mathcal{L}\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{F(s)}{s}.$$

Derivative theorem of the Laplace transform. If $f(t)$ is a Laplace transformable function with transform $F(s)$, and if $f(t)$ is continuous at $t = 0$, then

$$\mathcal{L}\{df(t)/dt\} = sF(s) - f(0).$$

The above properties, together with the linearity of the Laplace transform, allow the transform to be applied to any differential equation, or to any set of coupled equations that contain derivatives or integrals, such as those that result from circuit analysis. Note that initial conditions enter the analysis at the time the transform is applied, so the final solution automatically contains the component of the complete response due to initial conditions.

Since the one-sided Laplace transform does not consider negative time, and since initial conditions are automatically taken care of at the origin, Laplace transform techniques are extremely well suited to handle problems that contain switches that are turned on at the origin. Many analysts consider it the method of choice for circuit problems and differential equation problems in which a complete solution, including transient response terms, is required.

Finally, it should be noted that if the Laplace transform is applied to either the convolutional expression in Eq. (1) or to the differential equation in Eq. (2), the ratio of the transforms of the input and output can be found, that is, $H(s) = Y(s)/X(s)$, where $H(s) = \mathcal{L}\{h(t)\}$ is the transfer function of the system. The transfer function $H(s)$ completely characterizes a linear system, in much the same way as for the impulse response. In particular, stability and natural response behavior of the system can be determined directly from the poles and zeros of $H(s)$. Also, the frequency response is easily found by evaluating $H(s)$ at $s = j\Omega$. If a sinusoid of the form $x(t) = A\cos(\Omega_0 t + \phi)$ is applied to a linear, time-invariant, and causal system with transfer function $H(s)$, then the sinusoidal steady-state response will be $y_{ss}(t) = A|H(j\Omega_0)|\cos[\Omega_0 t + \phi + \arg H(j\Omega_0)]$.

C. Role of the Fourier Transform in Analog System Analysis

A close relative of the Laplace transform, which is also very useful in system analysis, is the Fourier transform, defined as

$$F(j\Omega) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{+\infty} f(t)e^{j\Omega t}\,dt \tag{4a}$$

$$x(t) = \mathcal{F}^{-1}\{F(j\Omega)\} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(j\Omega)e^{-j\Omega t}\,d\Omega, \tag{4b}$$

where the region of convergence is the entire $j\Omega$ axis whenever the transform exists. Note that the Fourier transform is identical to the two-sided Laplace transform with $s = j\Omega$. This implies that the Fourier transform of a causal system with impulse response $h(t)$ is identical to the frequency response of the system. Since the Fourier transform is two-sided, it contains information about behavior of the system for $t < 0$, and hence initial conditions do not arise in the same way as they do in applications of the Laplace transform. In general, analysts use the Fourier transform for sinusoidal steady-state analysis, although, theoretically, it also has the necessary properties to solve differential equations and obtain transient solutions.

Although they are closely related, it must be emphasized that the Laplace and Fourier transforms are quite different in nature, and hence they are both very useful in applications. Some of the differences are seen in Table II,

TABLE II Fourier Transforms of Some Common Functions[a]

Time functions, $f(t)$	Fourier transform, $F(j\omega)$
$k\,\delta(t)$	k
k	$2\pi k\delta(\omega)$
$\xi(t)$	$\pi\delta(\omega) + \dfrac{1}{j\omega}$
$\mathrm{sgn}(t)$	$2/j\omega$
$\cos\omega_0 t$	$\pi[\delta(\omega-\omega_0)+\delta(\omega+\omega_0)]$
$\sin\omega_0 t$	$j\pi[\delta(\omega+\omega_0)-\delta(\omega-\omega_0)]$
$e^{j\omega_0 t}$	$2\pi\delta(\omega-\omega_0)$
$t\xi(t)$	$j\pi\delta^{(1)}(\omega) - \dfrac{1}{\omega^2}$
$\sum_{k=-\infty}^{\infty}\delta(t-kT)$	$\omega_0\sum_{n=-\infty}^{\infty}\delta(\omega-n\omega_0), \quad \omega_0 = 2\pi/T$
$\sum_{n=-\infty}^{\infty}F_n e^{jn\omega_0 t}$	$2\pi\sum_{n=-\infty}^{\infty}F_n\delta(\omega-n\omega_0)$
$\dfrac{d^n\delta(t)}{dt^n}$	$(j\omega)^n$
$\lvert t\rvert$	$\dfrac{-2}{\omega^2}$
t^n	$2\pi j^n\dfrac{d^n\delta(\omega)}{d\omega^n}$

[a] Reprinted with permission from Gabel, R. A., and Roberts, R. A. (1980). "Signals and Linear Systems," 2nd ed., p. 326. Copyright © 1980 by John Wiley & Sons, Inc., New York.

which lists the Fourier transforms of some common functions, many of which we listed previously in Table I. Many of the Fourier transforms contain impulse functions that were not seen in the corresponding Laplace transforms. These impulse functions arise from the two-sided nature of the Fourier transform and often complicate solutions in simple problems. Note also that the Fourier transform of the sine and cosine functions are pure impulse functions, a result that coincides with our intuitive interpretation of frequency content of these signals.

III. ANALOG FILTERS REQUIRED IN A/D AND D/A CONVERSION

The traditional analog Fourier transform can be applied to an ideal sampled (analog) signal $s^*(t)$ as

$$\begin{aligned}\mathcal{F}\{s^*(t)\} &= \sum_{n=-\infty}^{+\infty} s(nT)\mathcal{F}\{\delta_a(t-nt)\} \\ &= \sum_{n=-\infty}^{+\infty} s_a(nT)e^{-jnT\omega} \\ &= \text{DTFT}\{s_a(nT)\}, \end{aligned} \tag{5}$$

where DTFT$\{\cdot\}$ denotes the familiar discrete-time Fourier transform of the discrete-time sequence $s_a(nT)$, which consists of samples of the analog signal $s_a(t)$. This verifies that the traditional Fourier transform of $s^*(t)$ is identical to the DTFT of $s(n) \equiv s_a(t)|_{t=nT}$, that is, the sequence $s(n)$ is derived from $s_a(t)$ by ideal sampling, and it shows that $s^*(t)$ and $s(n)$ are really different models of the same phenomenon, since their spectra (as computed with appropriate transforms) are identical. Suppose that $S_a(j\Omega)$ is the spectrum of $s_a(t)$, and $S(e^{i\omega})$ is the spectrum of $s(n)$. It can be shown that

$$S(e^{j\omega}) = \frac{1}{T}\sum_{r=-\infty}^{+\infty} S_a\left(j\left[\Omega - \frac{2\pi r}{T}\right]\right), \tag{6}$$

where $\omega = \Omega T$ is often referred to as the normalized digital frequency. Equation (6) shows that the DT spectrum is formed from a superposition of an infinite number of replicas of the analog signal, as illustrated in Fig. 2. As long as the sampling frequency $\Omega_s = 2\pi/T$ is chosen so that $\Omega_s > 2\Omega_B$, where Ω_B is the highest-frequency component contained in $s_a(t)$, then each period of $S(e^{i\Omega T})$ contains a perfect copy of $S_a(j\Omega)$, and $s_a(t)$ can be recovered exactly from $s(n)$ by ideal low-pass filtering. Sampling under these conditions is said to satisfy the Nyquist sampling criterion, since the sampling frequency exceeds the Nyquist rate $2\Omega_B$. If the sampling rate does not satisfy the Nyquist criterion, the adjacent periods of the analog spectrum will overlap, causing a distorted spectrum as illustrated in Fig. 2. This effect, called aliasing distortion, is rather serious because it cannot be corrected easily once it has occurred. In general, an analog signal should always be prefiltered with an analog low-pass filter prior to sampling so that aliasing distortion does not occur.

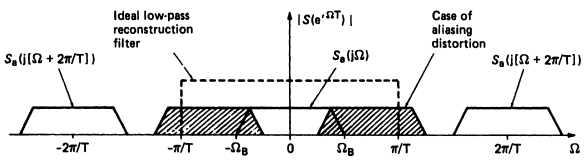


FIGURE 2 Relationship between the spectrum of an analog signal and the spectrum of the ideally sampled signal.

Figure 3 shows the frequency response of a fifth-order elliptic analog low-pass filter that meets industry standards for prefiltering voice signals, which are subsequently sampled at an 8-kHz sampling rate and transmitted digitally across telephone channels. The passband ripple is less than ± 0.01 dB from dc up to a frequency of 3.4 kHz (too small to be seen in Fig. 3); the stop-band rejection reaches at least -32.0 dB at 4.6 kHz and remains below this level throughout the stop band.

When the Nyquist sampling criterion is satisfied, it is always possible to reconstruct an analog signal from its samples according to

$$s_a(t) = \sum_{k=-\infty}^{+\infty} s_a(kT)\,\text{sinc}\left[\frac{\pi}{T}(t-kT)\right]. \tag{7}$$

This classical reconstruction formula can be derived by filtering $s^*(t)$ with an ideal low-pass analog filter with a bandwidth of $\Omega_B = \pi/T$, which removes the copies of the analog spectrum that were created by the sampling process (see Fig. 2). In general, exact reconstruction requires an infinite number of samples, although a good approximation can be obtained using a large but finite number of terms in the summation of Eq. (7).

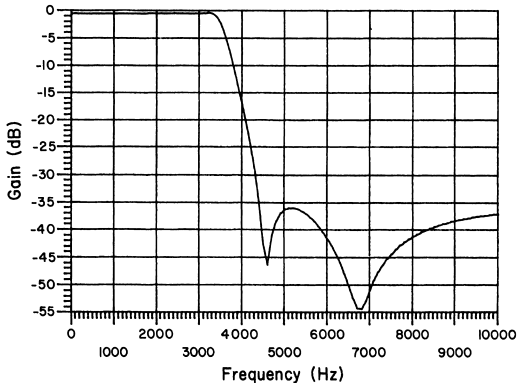


FIGURE 3 A fifth-order elliptic analog antialiasing filter used in the telecommunications industry with an 8-kHz sampling rate.

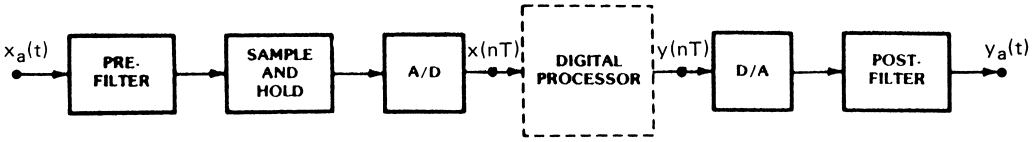


FIGURE 4 Analog pre- and postfilters required at the A/D and D/A interfaces.

Most practical systems use D/A converters for reconstruction, which results in a staircase approximation to the true analog signal, that is,

$$\hat{s}_a(t) = \sum_{k=-\infty}^{+\infty} s_a(kT)\{u(t-kT) - u[t-(k+1)T]\}. \quad (8)$$

It can be shown that $s_a(t)$ is obtained by filtering $s_a^*(t)$ with an analog filter whose frequency response is

$$H_a(j\Omega) = 2Te^{-j\Omega T/2}\text{sinc}(\Omega T/2). \quad (9)$$

The approximation $s_a(t)$ is said to contain $(\sin x)/x$ distortion, which occurs because $H_a(j\Omega)$ is not an ideal low-pass filter. $H_a(j\Omega)$ distorts the signal by causing a droop near the passband edge, as well as by passing high frequency distortion terms, which leak through the sidelobes of $H_a(j\Omega)$. Therefore a practical D/A converter is normally followed by an analog postfilter.

$$H_p(j\Omega) = \begin{cases} H_a^{-1}(j\Omega) & 0 \le |\Omega| \le \pi/T \\ 0 & \Omega \text{ otherwise,} \end{cases} \quad (10a)$$

which compensates for the distortion and produces the correct $s_a(t)$, that is, the correctly constructed analog output. Unfortunately, the postfilter $H_p(j\Omega)$ cannot be perfectly implemented, and therefore the actual reconstructed signal always contains, in practice, some distortion that arises from errors in approximating the ideal postfilter. Figure 4 shows a digital processor, complete with A/D and D/A converters, and the accompanying analog pre- and postfilters necessary for proper operation.

IV. ANALOG COMPUTING

Analog signal processing has many of its origins (even its name!) rooted in the historical development of the analog computer. Circuit designers found that with the aid of high-impedance, high-gain, vacuum-tube amplifiers they could design very high quality operational amplifiers by cascading numerous tube stages. These circuits inherited the name operational amplifiers because they could be used to realize accurately the mathematical operations of summation, scaling by constants, and integration. Given the capability to implement these operations, it was but a short time before engineers began to solve differential equations by building electrical circuits that would find solutions by electrical analogy. Simply stated, a circuit was built whose behavior is characterized by the same differential equation as the one to be solved. Then the physical behavior of the analogous circuits can be used to produce graphical solutions to many different excitations. Although the analog computer is very effective in producing real-time solutions to high-order linear differential equations, its real advantage is in the solution of nonlinear differential equations for which analytical techniques are sometimes unknown. As long as the nonlinearity can be modeled with an electronic circuit (such as a simple analog multiplier or saturation characteristic), the analog computer is quite effective in obtaining a solution.

Figure 5 shows the modern symbol for an operational amplifier. Although state-of-the-art op amps are now rather complicated solid-state circuits, they are characterized by a small number of rather simple properties at the input and output terminals. The standard op amp is a differential amplifier that has a very high voltage gain, typically of the order of 10,000 or more. It is characterized by a very high input impedance (it does not load circuits to which it is attached), a very low output impedance (it will drive any circuit that is attached to its output), and the differential gain is linear between the saturation limits of the amplifier. The high gain implies that as long as the amplifier is operating in its linear region, the difference in voltage between the positive and negative input terminals must be extremely small. This condition, referred to as a virtual short circuit at the input of a nonsaturated op amp, is used to simplify analysis of circuits that contain these types of amplifiers.

Figure 6 shows an op amp configured as a summer, where the input–output relation is given by

$$y_0(t) = -(R_F/R_1)x_1(t) - \cdots - (R_F/R_N)x_N(t). \quad (10b)$$

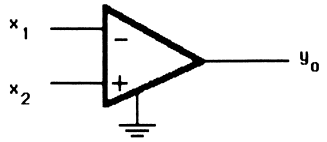


FIGURE 5 Modern symbol for an operational amplifier.

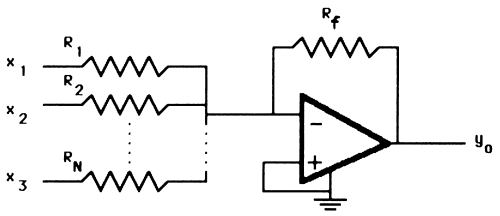


FIGURE 6 An op amp configured as a summer.

Similarly, Fig. 7 shows an op amp configured as an integrator, where the input–output relation is given by

$$y_0(t) = -\int_0^t [(1/R_1C)x_1(\tau) + \cdots + (1/R_NC)x_N(\tau)]\,d\tau. \tag{11}$$

By combining these basic circuits, it is possible to solve by analog simulation a wide class of linear or nonlinear differential equations.

A nonlinear analog problem will be illustrated by using the Bonhoeffer–Van der Pol model of a nerve membrane. The electrochemical response of a nerve cell is known to be a highly nonlinear function that displays an analog threshold phenomenon. The theoretical foundation for the Bonhoeffer–Van der Pol nerve model was established many years ago by pioneers in the field, who had derived equations for the ion transfers that take place in stimulated nerve membrane. From the ion-transfer equations, a four-dimensional model was developed in terms of a set of differential equations. The model is based on the classical nonlinear Van der Pol equation, which will be shown to form an interesting model of nonlinear nerve response.

The solution to the following second-order differential equation is a damped sinusoid of the form $x(t) = e^{-kt}(A\sin bt + B\cos bt)$:

$$d_y^2(t)/dt^2 + K\,dy(t)/dt + x(t) = 0, \left\{\begin{matrix}\text{initial}\\ \text{conditions}\end{matrix}\right\}. \tag{12}$$

If the damping constant k is replaced by a nonlinear term $c(x^2 - 1)$, the resulting equation, called the Van der Pol equation, is a nonlinear differential equation that describes the behavior of a large class of oscillators known as relaxation oscillators.

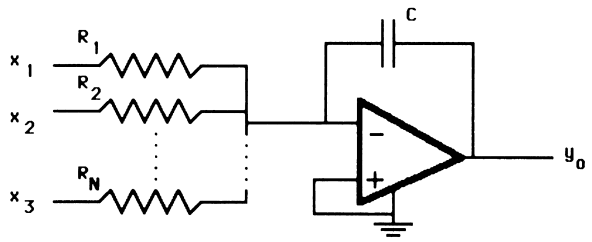


FIGURE 7 An op amp configured as a summing integrator.

$$d^2y(t)/dt^2 + c(x^2 - 1)\,dy(t)/dt + x(t) = 0, \left\{\begin{matrix}\text{initial}\\ \text{conditions}\end{matrix}\right\}. \tag{13}$$

By letting $y = (1/c)\,dx(t)/dt + x^3/3 - x$, Eq. (13) can be transformed into two-couple first-order differential equations, which can then be programmed on an analog computer:

$$dx(t)/dt = c(y + x - x^3/3) \tag{14a}$$

$$dy(t)/dt = -x/c \tag{14b}$$

In Eq. (13), if $c = 0$, then a plot of the phase plane x versus $dx(t)/dt$ is a circle. If c is a positive constant, during a portion of each cycle $x(t)$ will be small so that $c(x^2-1) < 0$, while during the remaining portion of the cycle $c(x^2-1) > 0$. Obviously, during part of the cycle the solution is decaying and the phase plot spirals toward the origin. During the other portion it grows and spirals away from the origin. An equilibrium condition is established between these opposing forces, and a nonlinear oscillatory condition is reached known as a limit cycle. In a limit cycle the phase plane trajectory appears as a distorted circle, the distortion arising from the nonlinearity of the system.

The Bonhoeffer–Van der Pol neuron model is obtained by modifying Eqs. (14) as

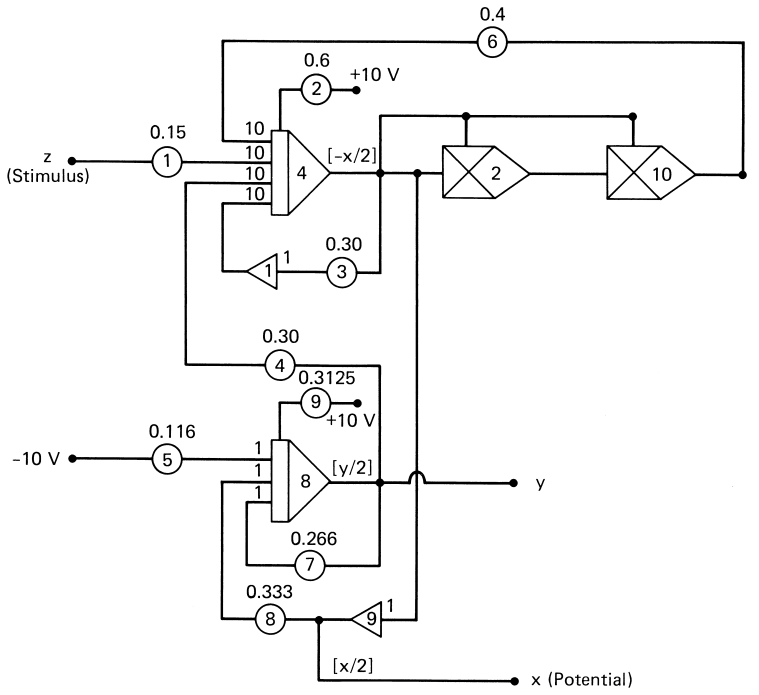
$$dx(t)/dt = c(y + a - x^3/3 + z) \tag{15a}$$

$$dy(t)/dt = -(x - a + by)/c \tag{15b}$$

The quantity z is a stimulus, while the quantity x is related to membrane potential.

Figure 8 shows the analog computer circuit diagram for this neural model as it was programmed on an analog computer with $a = 0.7$, $b = 0.8$, and $c = 3.0$. The stimulus z was chosen to be a pulse, ramp, and step for experimental purposes. Pulses of width 0.1 s were externally generated with a signal generator, with the height of the pulses carefully controlled by a potentiometer on the computer. The pulse generator was externally triggered by using a function switch and a positive reference voltage of 0.5 V. A function relay in series with a reference voltage served as a step input, while a ramp was created by integrating a reference voltage and using the comparator-hold circuits on the computer to stop the ramp and hold its final value. The pulse, step, and ramp stimuli were then used to investigate the important properties of this nonlinear model.

Limit-cycle behavior in the phase plane (x versus y) is shown in Fig. 9, with the different trajectories resulting from the application of different height pulses. The smaller trajectories about the resting point simulate subthreshold response. The neuron has not fired in this region, but there



Original equations

$\dot{x} = 3y + 3x - x^3 + 3z$

$\dot{y} = -0.33x - 0.266y + 0.233$

$x(0) = 12.0$

$y(0) = 6.25$

Scaled equations

$[\dot{x}/2] = 3[y/2] + 3[x/2] - 4[x^3/8] + 1.5z$

$[\dot{y}/2] = -0.33[x/2] - 0.266[y/2] + 0.116$

$[x/2(0)] = 6.0$

$[y/2(0)] = 3.125$

FIGURE 8 Analog computer diagram for the Bonhoeffer–Van der Pol neuron model.

has been a disturbance of potentials due to limited ion exchange. The "enhanced" and "depressed" regions indicate areas on the phase plane where, if the model is in one of these states when a stimulus is applied, the response is correspondingly affected.

Superthreshold neural response is simulated by the larger trajectories, along which regions have been labeled to correspond to physiological nerve states. One interesting feature of this model is that it is truly an analog model that exhibits threshold-type behavior. All states within the

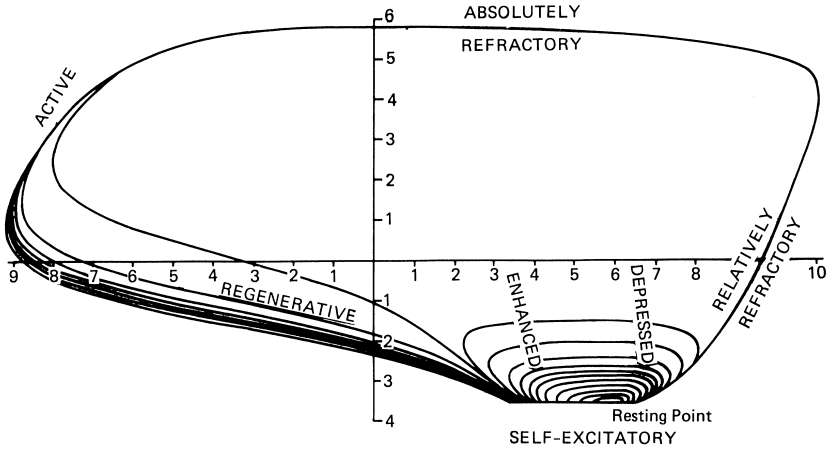


FIGURE 9 Limit cycle behavior of the Bonhoeffer–Van der Pol neuron model in the phase plane (x versus y).

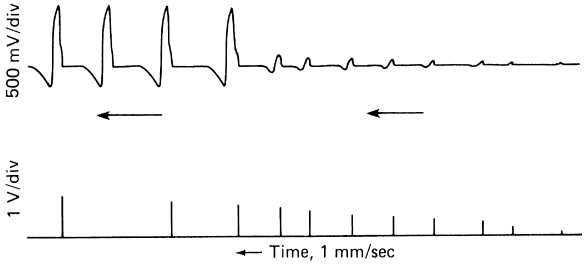


FIGURE 10 Illustration of threshold behavior.

phase plane can be reached, but the central region is nonattractive and is rarely traversed. This grouping of responses into large and small trajectories, called a quasi-threshold phenomenon in the literature, seems to describe the physical nerve behavior better than a digital description.

The time-domain operation of the Bonhoeffer–Van der Pol model is illustrated in Figs. 10–13. These figures illustrate many interesting properties of the model, including threshold behavior (Fig. 10), firing by rapid subthreshold excitation (Fig. 11), response to stimuli in the refractory state (Fig. 12), and continuous firing caused by persistent stimuli (Fig. 13). It is apparent that many of these interesting properties demonstrated by analog computer simulation are a result of the nonlinear properties of the underlying Van der Pol model.

V. CHARACTERISTICS OF ANALOG VLSI CIRCUITS

Over time there has been an increasing emphasis on the development of very large scale integrated (VLSI) circuits for use in analog signal processing, as opposed to the more traditional discrete types of circuits constructed from collections of many small scale or medium scale integrated (SSI or MSI) circuits. The VLSI circuits are physically smaller, require less silicon area, consume less power, are more reliable, and result in reduced costs in general. However, VLSI circuits are much more difficult to design and test due to the large number of components on the chip and the limited number of terminals that can be brought out to the external environment for testing. Computer-aided design, layout, and testing are very important for VLSI circuits because of the complexities that result from their size.

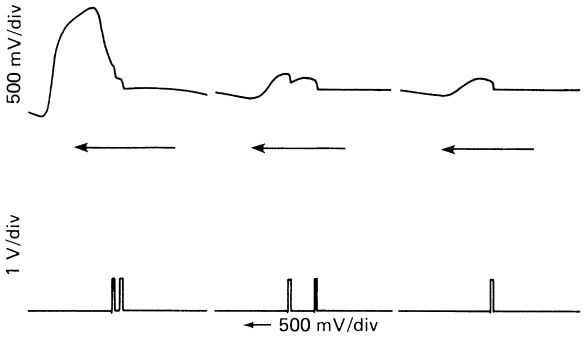


FIGURE 11 Response to subthreshold stimuli.

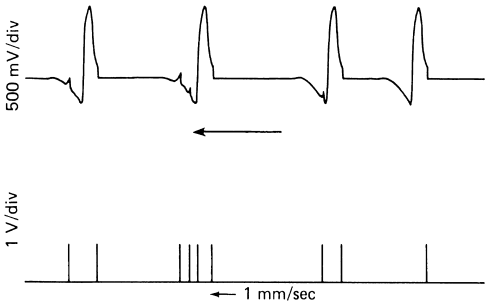


FIGURE 12 Response to stimuli during refractory period.

This section reviews analog and digital VLSI. Figure 14 shows an overview of the design process for a generic integrated circuit (IC), as representative of the current state of the art in industry. The overall design process is a complicated interaction between design, simulation, and fabrication. Since the overall process is tedious and expensive, the use of computer programs for design, analysis, and testing is extremely important. Table III summarizes the key similarities and differences between analog and digital design. In general, it is fair to conclude that analog design is more difficult and expensive than is digital. It is not as well suited for computer-aided design (CAD) tools, and it is not able to take as full advantage of VLSI capabilities as it is for digital design. Tables IV–VI summarize the essential characteristics of the three major integrated circuit technologies that are in use today: (1) bipolar, (2) complementary metal–oxide–silicon (CMOS), and (3) gallium arsenide (GaAs). Both bipolar and CMOS are the workhorses of the integrated-circuit industry today, while GaAs seems to hold a promise for the future, due to its potential for wide-band analog signal processing capabilities. This is more easily seen in Fig. 15, which

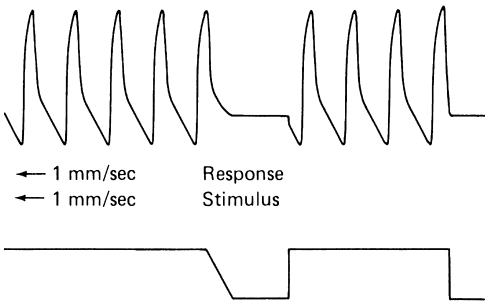


FIGURE 13 Continuous firing due to persistent stimuli.

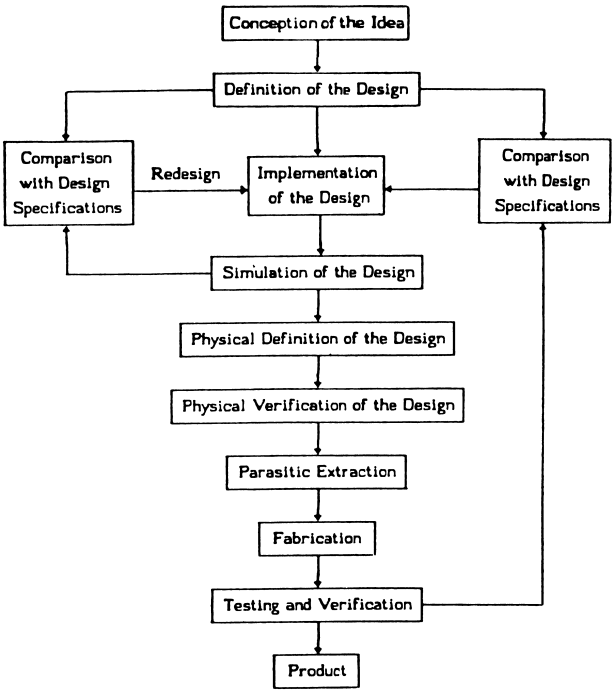


FIGURE 14 Overview of the IC design process. [From Allen, P. E. (1986). Analog VLSI design tutorial. (1986 Int. Symp. Circuits and Systems, Presymp. Workshop Analog and Digital VLSI, San Jose, California, May 1986.)]

compares approximate signal bandwidths for many different IC technologies. Note that GaAs (analog) technology lies at the upper extremes of signal processing bandwidths, in the same region as surface acoustic-wave devices and optical processors. Another interesting feature that can be seen in Fig. 15 is that although the analog and digital categories overlap in the regions from 100 Hz to 10 MHz, digital is not able to compete well in the upper range of 10 MHz to 100 GHz. Therefore, in spite of the favorable position occupied by digital VLSI in Table III, in the foreseeable future, analog VLSI will be the choice of circuit designers for signal processing in the very high frequency ranges.

TABLE III Analog versus Digital Design: Important Characteristics[a]

Analog design	Digital design
Signals have a continuum of values for amplitude and time	Signals have only two states
Irregular blocks	Regular blocks
Customized	Standardized
Components have a continuum of values	Components have fixed values
Requires precise modeling capability	Optimizes VLSI technology
Difficult to use with CAD	Amenable to CAD tools
Performance optimized	Programmable by software
Designed at the circuit level	Designed at the system level

[a] From Allen, P. E. (1986). Analog VLSI design tutorial. (1986 Int. Symp. Circuits and Systems, Presymp. Workshop Analog and Digital VLSI, San Jose, California, May 1986.)

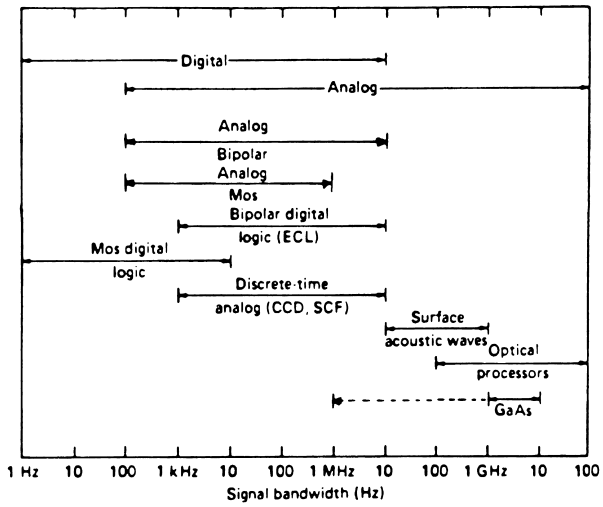


FIGURE 15 IC technology versus signal bandwidth. ["Design Methodology for VLSI Circuits" (Sidhoff and Nordhoff, eds.). Nato Advanced Study Institute on Design Methodologies for VLSI Circuits, Louvain La Neuve, 1982.]

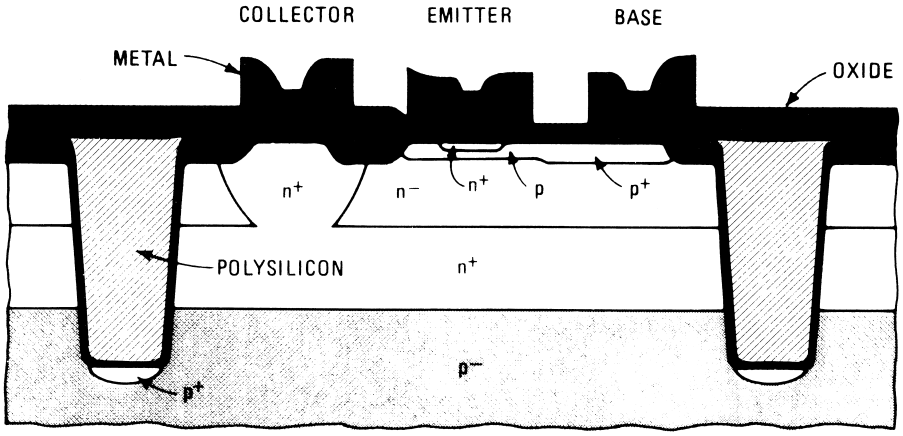
VI. SWITCHED-CAPACITOR CIRCUITS

In the mid-1970s switched-capacitor (SC) circuits became popular with circuit theorists and IC technologists when it was discovered that SC techniques could be used in high-precision monolithic filters. Driven by expectations of high economic payoff, SC circuits became a popular topic at workshops and conferences and in trade magazines and scientific journals, as well as throughout the entire IC industry. The explosion of activity in SC circuits tended to obscure the years of "awakening." Similarly, the frenzy of commercial development, with its focus on new products, tended to overshadow some of the important theoretical advances that were occurring at this time. This section is not meant to be a comprehensive historical survey, but rather an attempt to identify some of the important developments that marked key events in the evolution of the subject.

Prior to the discovery of SC circuits, the high-precision monolithic frequency-selective circuit was a tantalizing concept that remained elusive in practice. This was partly due to the fact that ICs cannot contain high-Q inductors, since the Q of an inductor decreases with decreasing size.

TABLE IV Summary of Bipolar Integrated Circuit Technology[a]

1–3 μm, Trench isolation techniques ± 3 to ± 5 V power supplies
Up to four levels of metal
Bipolar structures and processing becomes simpler as the device is scaled down
Capacitors are not easily implemented
Low noise
High transconductance
Fast switching
Stabilized technology



Fujitsu's trenched oxide and polysilicon isolation BJT.

[a] From Allen, P. E. (1986). Analog VLSI design tutorial. (1986 Int. Symp. Circuits and Systems, Presymp. Workshop Analog and Digital VLSI, San Jose, California, May 1986.)

Integrated resistors are nonlinear, occupy large silicon areas, and their tolerances and temperature coefficients do not closely track variances in the capacitors on the same chip. This results in RC time constants that cannot be controlled accurately and produces monolithic circuits with large variances in the frequency characteristics. In contrast, highly accurate capacitors can be realized quite easily in metal–oxide semiconductor (MOS) technology. Since different capacitors on the same chip track each other with variations in processing, capacitor ratios can be realized very accurately. Switched-capacitor circuits are designed to depend only on capacitor ratios and thereby take advantage of the property to realize high-precision circuits.

After some initial publications of SC circuit design concepts circa 1977, a great amount of research began in the design and analysis of more general classes of SC circuits. Some of the initial enthusiasm decreased as the subject evolved into maturity. This observation is supported by the data presented in Fig. 16, which show the number of papers presented at the International Symposium on Circuits and Systems from 1977 to 1983 on the subject of SC circuits. This number increased rapidly from 1977 to 1981, peaked in 1982, and then fell significantly in 1983. This does not mean that the subject decreased in importance, or that all important problems were solved by 1983, but rather that the initial excitement had passed, the easy problems had been solved, practical limits had been identified, and circuit designers were beginning to face much more difficult problems.

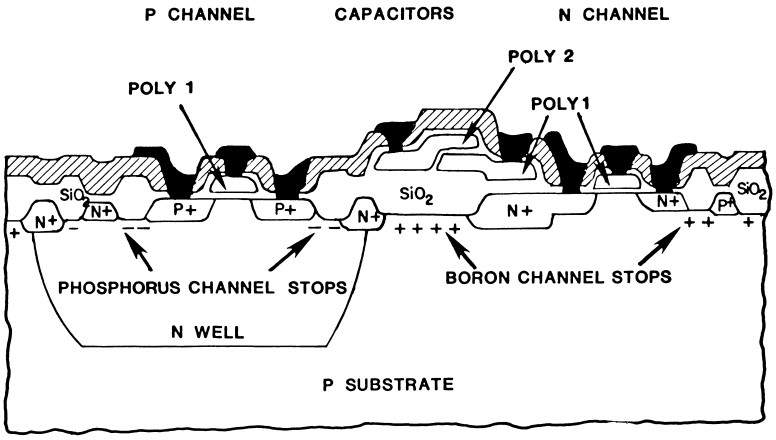
A. Historical Perspectives on SC Circuits

Early researchers recognized that linear analog systems that contain switches represent an important class of linear time-varying (LTV) systems for which there exists a theory of analysis and design. If the switches are operated periodically, the system becomes linear periodically time-varying (LPTV), which is the closest relative to ideal linear time-invarant (LTI) systems. The inclusion of switches in an otherwise LTI network creates a more general structure with enhanced capabilities, although the mathematics required for analysis and design become more complicated.

Much of the early work on switched systems was done in Western Europe for telephone communication systems. In notable works on the theory of resonant transfer structures, capacitances are made to exchange charges periodically through appropriate interconnections with resonant circuits. Although the original resonant transfer circuits

TABLE V Summary of CMOS Integrated Circuit Technology[a]

1–3 μm, Single well
± 3 to ± 5 V Power supplies
Double polysilicon
Several layers of metal
Substrate BJT
Good capacitor
Good switch
Low power
Problems with scaling—second-order effects, electron effects, etc.
Mature but still growing technology

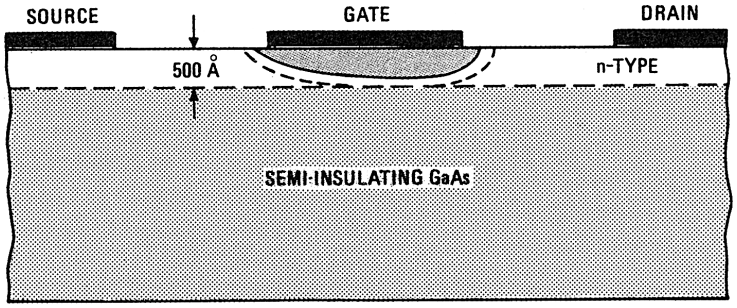


An example of state-of-the-art CMOS technology.

[a] From Allen, P. E. (1986). Analog VLSI design tutorial. (1986 Int. Symp. Circuits and Systems, Presymp. Workshop Analog and Digital VLSI, San Jose, California, May 1986.)

TABLE VI Summary of Gallium Arsenide Integrated Circuit Technology[a]

0.5–3 μm, MESFET
± 4 to ± 6 V Power supplies
$f_T \simeq 10 f_T$ (CMOS)
Poor hole mobility, only *n*-type devices
Wider temperature range
Not yet efficient for VLSI—conductor crossing, models, high power dissipation, etc.
Immature technology
Wide-band analog signal processing (op amp: 60 dB and 1000 MHz unity-gain bandwidth



Example of GaAs MESFET technology.

[a] From Allen, P. E. (1986). Analog VLSI design tutorial. (1986 Int. Symp. Circuits and Systems, Presymp. Workshop Analog and Digital VLSI, San Jose, California, May 1986.)

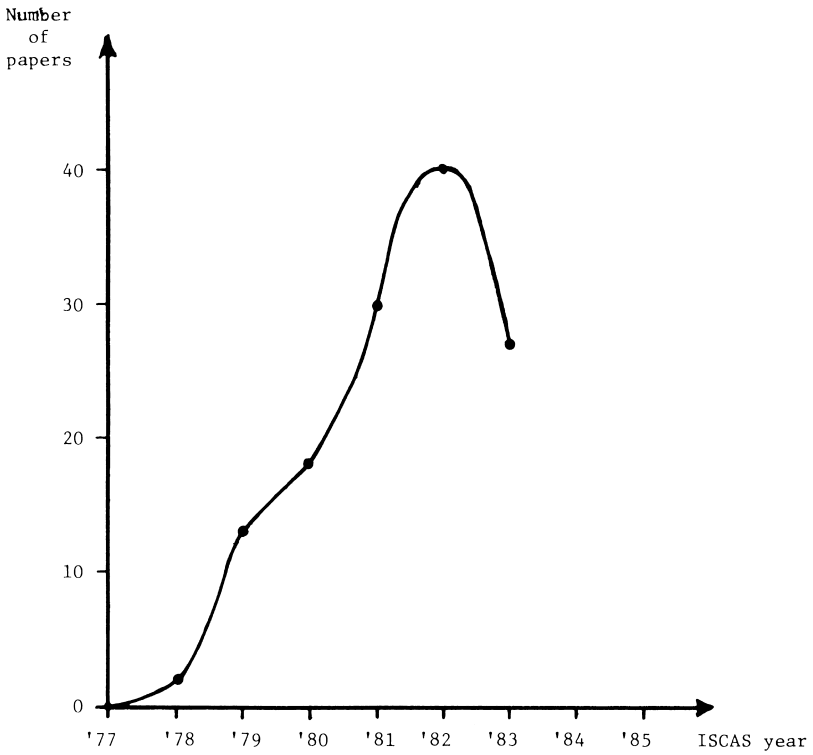


FIGURE 16 Number of papers on SC circuits presented at the IEEE International Symposium on Circuits and Systems from 1977 to 1983.

were largely replaced later by digital pulse-coded modulation (PCM) techniques, they provided a catalyst for the development of a more general theory of LPTV systems that become useful for the beginning studies of SC circuit analysis.

A second important development in the early years was the N-path filter structure. In its basic form this structure consists of N LTI systems arranged in parallel and switched into the system sequentially so that only one subsystem operates during a specified time interval. A number of authors used the N-path structure as a generic model for SC circuits, where each subsystem represents the SC circuit during one phase of the switching sequence. Also, the N-path model has been useful in SC circuit analysis programs because it decomposes the analysis into a sequence of LTI subsystem analyses in which each subnetwork provides initial conditions for the subnetwork that follows in the switching sequence.

The flurry of SC activity began in the mid-1970s. By 1967 the idea had already been established that a filter characteristic could be specified entirely by capacitor ratios. It now seems clear that the burst of new activity in the mid-1970s was not a result of new advancements in either circuit theory or IC technology alone but rather was the result of coupling these two areas in an effective way.

The following rush of excitement was, to a great extent, led by IC technologists, that is, by researchers with the experience, facilities, and initiative to fabricate and test the first generation of experimental SC circuits. The group at the University of California at Berkeley reported success with fabricating low-pass antialiasing filters to be used as transmit prefilters and receive postfilters in telephone channel coder–decoders (CODECs). The first SC filter designs were derived from active filters in the state variable configuration by means of resistor replacement. (Note the analog computer influence in this development!) For example, the second-order filter shown in Fig. 17 is a state variable biquad in which the two resistors were replaced by switched capacitors $\alpha_1 C$ and $\alpha_2 C$. The first resistor is replaced by a "toggled" switched capacitor, whereas the second was replaced by a "parallel" switched capacitor. It was determined experimentally that the use of the toggled SC in the input path created a sharper rolloff than it would have had if a parallel SC been used in that branch also. However, in the early days of 1978, it was not clear how to choose the arrangement of switches to produce the best filter. This started a new movement to analyze SC circuits as discrete-time (but not digital) filters rather than as approximate active filters. Although they were not sophisticated filters to the circuit theorist, circuits of this type were fabricated at Berkeley, representing a first

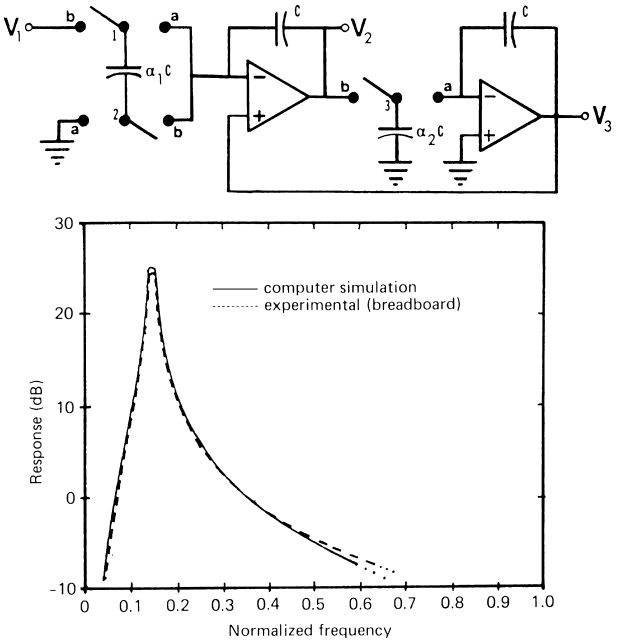


FIGURE 17 State-variable SC biquad by resistor replacement.

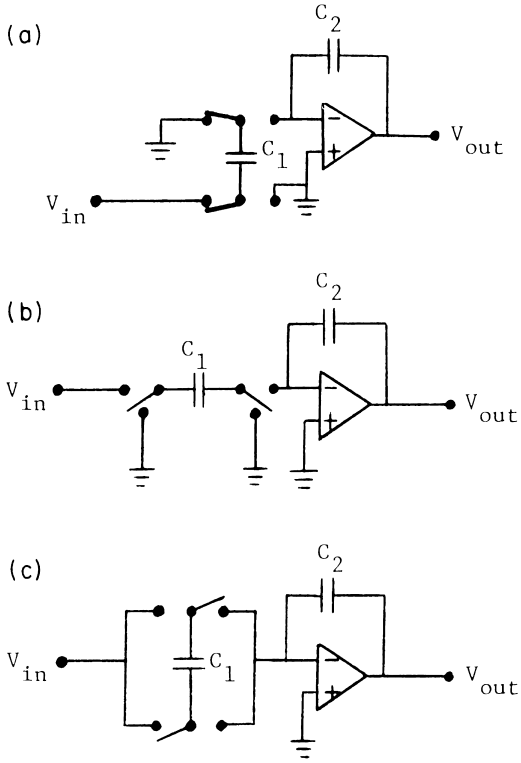


FIGURE 18 SC integrators: (a) forward Euler integrator, (b) backward Euler integrator, and (c) trapezoidal integrator.

generation of circuits to demonstrate the feasibility of the SC concept.

When it became clear that resistor replacement techniques are inaccurate for switching rates near the Nyquist limit, designers began to characterize charge transfer with difference equations, derive transfer functions in the Z domain, and apply the theory of discrete-time systems to analyze performance. It was learned that if the output voltage is sampled on the "a" switch phase, then the network shown in Fig. 18a behaves as a forward Euler integrator, the network of Fig. 18b is a backward Euler integrator, and that of Fig. 18c as a trapezoidal integrator. Very high quality SC filters can be derived from the "leapfrog" active filter structure, which is a configuration of a doubly terminated LC ladder using these types of integrator replacements. The switch phasing between integrators in ladder networks also greatly affects performance. By simply alternating the phases between integrators, the ladder was converted from an Euler-sampled data design to a lossless discrete integrator (LDI) structure that is well known in the design of digital ladder filters. This was an important result because it showed that the use of discrete-time system theory can in certain circumstances lead to greatly improved performance without any increase in cost.

In 1978 a discovery was made that initially excited circuit theorists a great deal but that later led to considerable frustration because of practical limitations. It was discovered that a toggled switched capacitor, similar to the one used to replace the resistor R_1 in the state variable biquad of Fig. 17, transforms the analog structure into an SC structure according to a bilinear transformation that is well known as a design technique for digital filters. If each phase of the biphase clock has period T, the effective sampling rate becomes $2/T$ because the toggled switched capacitor has identical behavior on each phase. At this time, many circuit theorists thought that SC filter design theory was complete with this discovery, in that a good SC filter could be obtained from an active RC filter by simply substituting a toggled switched capacitor for each resistor. However, there are practical limitations that prevent this approach from working well in practice.

In MOS transistors there normally is a rather large nonlinear parasitic capacitance C_s between the bottom plate and the substrate, as well as a considerably smaller parasitic capacitance C_p between the top plate and the substrate. The value of C_s can be as much as 20% of the fabricated capacitor and is rather difficult to predict or compensate. The value of C_p is normally an order of magnitude smaller, although its presence can disturb a sensitive design significantly. It is important for SC circuits to be designed with the bottom plate of all capacitors grounded or switched between a voltage source and ground, so that the excess charge accumulated by the parasitic during the charging phase is harmlessly discharged to ground and does not affect the characteristics of the circuit. It is clear

that a toggled switched capacitor cannot be made parasitic-insensitive in general, since the top and bottom plates must perform the same function on alternate phases. Despite efforts to compensate for parasitics in the bilinear switched capacitor, it was largely rejected by IC technologists as impractical because it led to too much uncontrolled variation in the characteristics of the final SC circuits.

Note that the forward and backward Euler integrators of Fig. 18 are insensitive with respect to both the top plate and bottom plate parasitics. For this reason these types of integrators became more successful than the bilinear integrator, and were subsequently used in many design methods that were eventually published. In 1979, a set of practical considerations for MOS implementation of SC circuits was published:

1. Switched-capacitor "resistors" should not close an op amp feedback path.
2. There should be no floating nodes.
3. At least one plate of every capacitor should be connected to a voltage source or switched between voltage sources.
4. The noninverting op amp input should be kept at a constant voltage source (usually grounded) or switched between voltage sources.

Although items mentioned in this list should be treated as guidelines rather than as strict rules, they seem to have withstood the test of time in the sense that most IC technologists tend to reject designs that violate these conditions.

In 1979 a design approach was published that incorporated these rules of thumb a priori and represented a considerable change of philosophy from the resistor and integrator replacement techniques that dominated earlier thinking. A network topology was chosen for a general biquad that incorporated the above guidelines. Biphase switching was selected that eliminates continuous feed-through paths and design equations were derived that express capacitor ratios in terms of desired pole locations. This represents one of the most successful attempts at designing an SC biquad directly from $H(z)$. The approach has the added advantage that the network topology is predetermined so that the IC fabrication process can be standardized to a considerable degree.

For example, the circuit of Fig. 19 is a second-order low-pass notch filter designed by this technique. The notch was placed at 1.8 kHz, as seen in the frequency response of Fig. 20. Note that the frequency analysis shows a somewhat different magnitude response in the notch on the two phases, although one can easily determine by inspection that the magnitude response of this biquad is theoretically the same on each phase. Since it was suspected that the different notch responses were due to computational inaccuracy in the frequency-analysis routine, a sensitivity analysis was performed for the capacitors, the most sensitive of which are shown in Table VII. The sensitivity is two orders of magnitude larger in the notch in comparison with other frequencies in the pass region. It is interesting to note that the notch could not be obtained experimentally with average-quality discrete components;

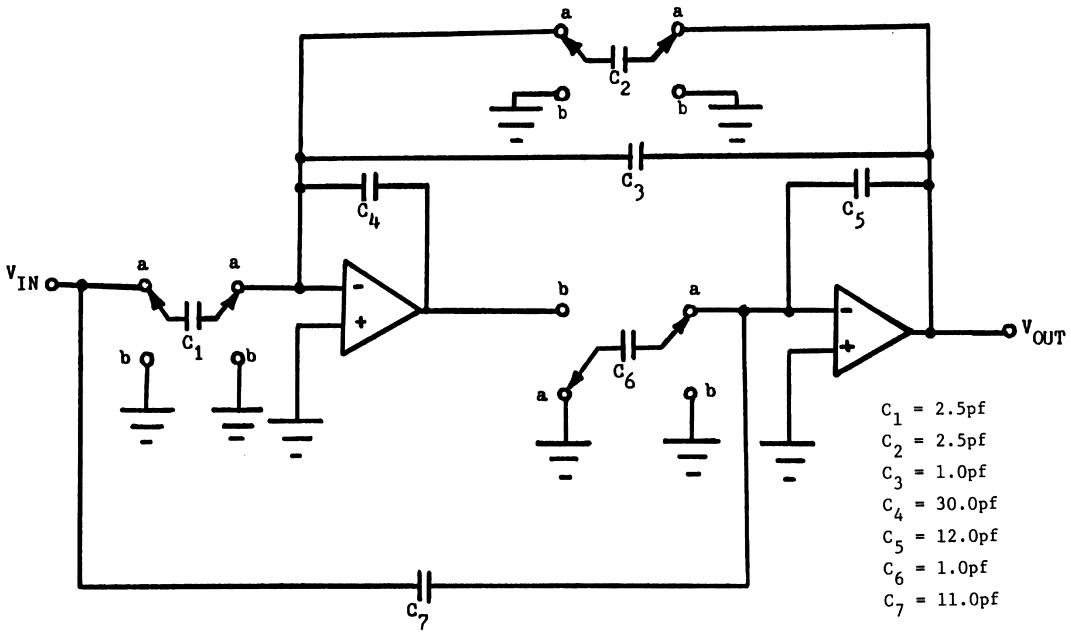


FIGURE 19 A second-order SC notch filter.

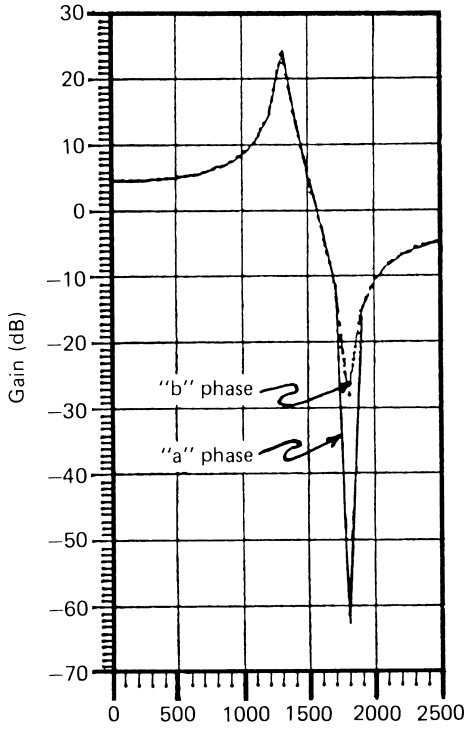


FIGURE 20 Frequency response for the notch filter of Fig. 19.

it is not clear whether the integrated form of this network would achieve satisfactory rejection at the notch frequency.

There have been many more attempts to establish more general theories of analysis and design of SC circuits. Also, there have been numerous studies of sensitivity noise and op amp finite-gain bandwidth effects. In general there are remarkable similarities in the properties and performance merits between active filters and their corresponding SC realizations. In particular, the active structures and their SC relatives track closely with respect to sensitivity and noise performance.

TABLE VII Sensitivities for the Most Critical Capacitors in the Notch Filter of Fig. 19

Frequency (Hz)	Magnitude sensitivity		
	C_1	C_6	C_7
1500.0	0.3280E+01	−0.1204E+01	−0.2280E+01
1550.0	0.3879E+01	−0.1942E+01	−0.2879E+01
1600.0	0.4782E+01	−0.3521E+01	−0.3781E+01
1650.0	0.6291E+01	−0.7217E+01	−0.5290E+01
1700.0	0.9322E+01	−0.9870E+01	−0.8328E+01
1750.0	0.1851E+02	−0.3058E+02	−0.1752E+02
1800.0	−0.5189E+03	−0.5163E+03	−0.5201E+03
1850.0	−0.1739E+02	−0.1229E+02	−0.1838E+02
1900.0	−0.8669E+01	−0.4809E+01	−0.9668E+01
1950.0	−0.5721E+01	−0.2641E+01	−0.6720E+01
2000.0	−0.4241E+01	−0.1692E+01	−0.5241E+01

B. Computer-Aided Analysis of SC Circuits

In the late 1970s there appeared to be two approaches to the computer-aided analysis of SC circuits; these were, in retrospect, temporary solutions. The first was the direct use of the well-known SPICE program, although users quickly discovered that this approach is very expensive in terms of computer time and really generated much more information about transients between switching states than was required. The second approach was to model an SC circuit as a digital filter with adders, multipliers, and delay registers, and then use a digital filter analysis program such as DINAP for time, frequency, and noise analyses. While this approach was effective for simple circuits with two-phase switching, the modeling proved to be too time-consuming for more complicated circuits, especially those with multiple clock phases.

Table VIII gives a good perspective of the state of the art of SC CAD programs in 1981. There clearly was an evolution from programs that dealt with specific topology and two-phase switching, to programs that were completely general with regard to topology, number of switching phases, phase intervals, and input–output (I–O) coupling. It seems that no one program has achieved the type of widespread acceptance that was extended to SPICE previously by the scientific community, although there are several that have been used considerably outside the institutions where they were developed. SWITCAP, developed at Columbia University, is a user-oriented program that allows general network topology, general N-phase switching sequences, and continuous I–O coupling. It performs time-domain and frequency-domain analyses while permitting continuous signal paths. A particular feature included in this program is a compaction algorithm that reduces the dimension of the system equations prior to executing the sparse matrix solution. The equations are essentially nodal equations formed from appropriately defined closed surfaces that represent conservation of charge relations. DIANA, developed at Katholieke Universiteit Leuven, is a well-known CAD package that was modified to admit SC circuits in addition to its original mixed-mode simulation capabilities. DIANA has a wide range of capabilities that lean heavily on the routines that previously existed for sparse matrix solution, sensitivity analysis, etc. DIANA is the only program that was available on a commercial basis at the time of this study. Equation formulation in DIANA is based on an adapted form of modified nodal analysis and the sensitivity analysis takes advantage of adjoint techniques. SCAPN, developed at the University of Illinois, is a third-generation

TABLE VIII CAD Programs for SC Circuits[a]

Source of the program	A. Ganesan Bell Labs U.S.A.	H. De Man *et al.* Kath. Univ. Leuven Belgium	R. Schweer *et al.* Univ. Dortmund Germany	U. Kleine *et al.* Univ. Dortmund Germany	H. Gutsche Siemens Germany	A. Knob and R. Dessoulavy EPFL Switzerland
Name of the program	**BSCAP**	**DIANA**	**DOSCA**	**DONES**		
Features						
Time-domain analysis	Yes	Yes	Yes		Yes	
Frequency-domain analysis	Yes	Yes	Yes		Yes	Yes
Aliasing effect	Yes	Yes				
Sensitivities		Yes	Yes			
Group delays	Yes	Yes	Yes			
Noise analysis		Yes	Yes			
Nonlinearities		Yes	Yes			Yes
Resistor or op amp poles	Yes	Yes	Yes		Yes	Yes
More than 2 phases	Yes	Yes	Yes		Yes	
Continuous I–O coupling	Yes	Yes				
Topological restrictions			Yes			Some
Design synthesis				Yes		
Optimization			Yes	Yes		
Mixed mode and top down design		Yes				
Symbolic transfer functions						
Multiple clock frequencies						
User's manual available			Yes			
Availability	No	Yes	Yes	No	No	Yes Implement in SPI, CE, ASTP, etc.
Performance						
Accuracy problems					Small and large resistance	
Size of network					Degree <16	
Speed						

continues

SC program (following SCAPI and SCAPII of Table VIII) that also accommodates SC networks with general topology and arbitrary switching sequences. It is a stand-alone program like SWITCAP in the sense that it was developed specifically for SC circuits and does not exist as a subprogram of a more general software package, as in DIANA. SCAPN uses modified nodal analysis directly in the Z domain, as well as sparse-matrix storage, adjoint techniques for sensitivity analysis, and a special pre-LU factorization algorithm that conserves computation during frequency updates of the LU factors. SCAPN factors a frequency-domain nonlinear distortion-analysis routine for nonlinear capacitors and op amp gains, although in its present form it does not permit continuous I–O coupling.

There are many other analysis programs and problem formulation philosophies described in the literature, and it is regrettable that they cannot all be discussed here. In the earlier sections of this article emphasis was placed on approaches that focus on the discrete-time nature of the SC circuit. There is also a different school of thought represented in the literature, in which the problem of modeling the discrete-time behavior of the circuit is approached by developing analog equivalent circuits so that the subsequent analysis can be done with existing analysis programs. For example, this approach has been popular within Bell Laboratories, where the circuit analysis is ultimately carried out using the well-known analog analysis program CAPECOD. It seems that both approaches to SC computer-aided circuit analysis result in valid simulation of circuit behavior, so it is likely that both will be continued in future applications.

TABLE VIII CAD Programs for SC Circuits[a] (*continued*)

E. Simonyi *et al.* TKI Budapest Hungary	C. Lee and W. Jenkins Univ. of Illinois U.S.A.	C. Lee and W. Jenkins Univ. of Illinois U.S.A.	R. Plodeck *et al.* ETH-Zürich Switzerland	M. Bon, CNET France A. Konczykowska T. Univ. Warsaw, Poland	Y. Tsividis *et al.* Columbia Univ. U.S.A.	T. Rübner-Petersen Lyngby Tech. Univ. Denmark
PRSC	**SCAPI**	**SCAPII**	**SCANAL**	**SCYMBAL**	**SWITCAP**	**SCNET**
	Yes			Yes	Yes	
Yes	Yes	Yes	Yes	Yes	Yes	Yes
			Yes		Yes	
	Yes	Yes		Yes		Yes
				Yes		
						Yes
Yes			Soon	Yes	Yes	Yes
		Yes				
	Some	Some				
Yes						
Yes						
			Yes	Yes		
					Yes	
Yes	Yes	Yes	Yes	Yes	Yes	Soon
Yes	Yes	Yes	Yes	Yes	Yes	Yes
On-desk-top computer						
		Some at zero frequency				
Small				Small		
	Fast	Slower				

[a] Reprinted with permission from Vandewalle, J. (1981). *Proc. 1981 Eur. Conf. Circuit Theory and Design, The Hague, August 1981*, p. 1072. Copyright © 1981 by North-Holland, Amsterdam.

C. Switched-Capacitor Pre- and Postfilters: A Case Study

One successful application of switched-capacitor circuit technology is in the realization of the analog pre- and postfilters required for digital voice transmission in modern telephone channels. The A/D and D/A conversion, as well as some additional signal conditioning (companding, etc.), is done with a type of IC referred to as a CODEC (for coder–decoder). One IC contains the transmit filter, to handle transmission in the forward direction, the receive filters, to handle the returned voice signal, and any other circuits required for signal conditioning. The industry standard for the digital sampling rate is 8 kHz, which is designed to prevent aliasing on a 4-kHz telephone channel. Prior to SC circuit technology, the transmit and receive filtering was done with an active filter that was implemented by hybrid fabrication techniques, which added considerable cost to the system. With switched-capacitor circuit techniques, the goal of integrating the transmit and receive filters together with the CODEC, while meeting the rather strict system specifications, became reachable. The philosophy is to switch the SC filters at a very high rate (usually 256 kHz) relative to the 8-kHz sampling rate of the CODEC so that sampled data effects in the SC circuits are negligible, and for all intents and purposes the SC pre- and postfilters function as analog filters. The transmit filter must have a low-pass characteristic

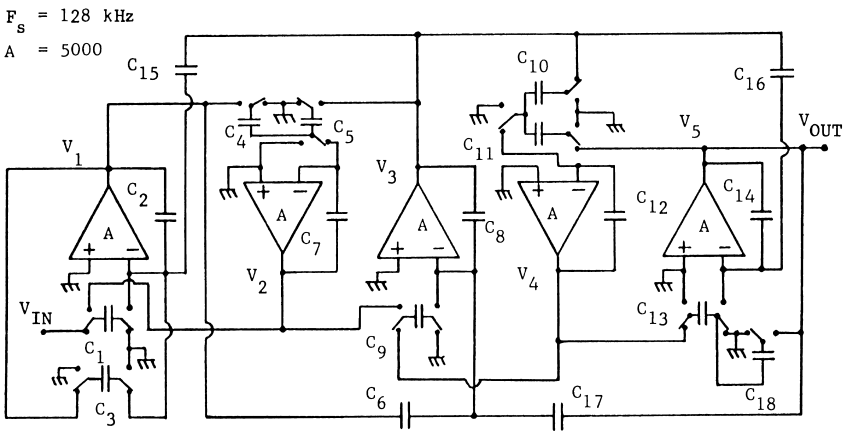


FIGURE 21 A fifth-order elliptic low-pass state-variable ladder realizatio.

within a tolerance of 0 ± 0.01 dB in the pass-band up to a ripple cutoff frequency of 3.4 kHz. At 4.6 kHz the filter rejection must be at least -32.0 dB and the response must remain below this level for all frequencies above 4.6 kHz. In addition, the transmit filter must have (at least) a -20 dB notch at 50 to 60 Hz to reject 60 (50) cycle noise that is usually present in the environment In the receive channel, a postfilter is required following the D/A converter to smooth the staircase waveform that results from the reconstruction process (see Section III for a discussion of the theory). The receive filter is a low-pass filter with a cutoff at 3.4 kHz, which is essentially the same as the antialiasing transmit filter, although there is an additional requirement for a gentle 1.7- to 1.9-dB gain in the pass-band to compensate for the $(\sin x)/x$ distortion characteristic of the D/A reconstruction process. The example designs shown here are for the idealized lowpass filter alone; that is, they do not include the transmit notch and receive $(\sin x)/x$ compensation. It is well known in the telecommunications industry that systems specifications can be met with a fifth-order elliptic design for the ideal lowpass filters.

The three SC ladder structures shown in Figs. 21–23 are all capable of realizing the required fifth-order elliptic low-pass transmit and receive filters just described. The circuit of Fig. 21 is a variation of the state-variable ladder derived by an LDI (lossless discrete integrator) transformation, although certain switches have been rearranged to guarantee a parasitic-insensitive circuit. Figure 22 shows a second design based on inductor simulation,

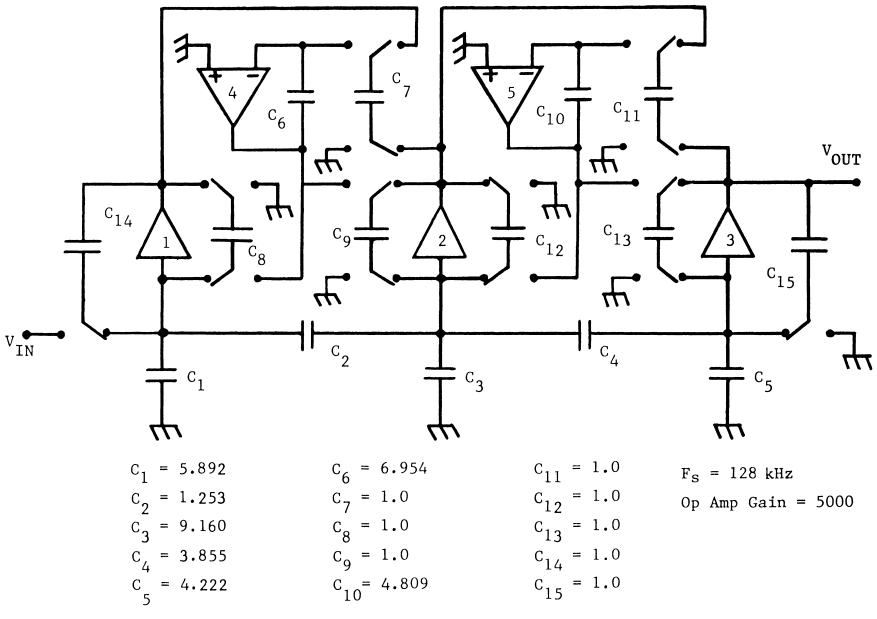


FIGURE 22 A fifth-order elliptic low-pass ladder realization based on inductor simulation.

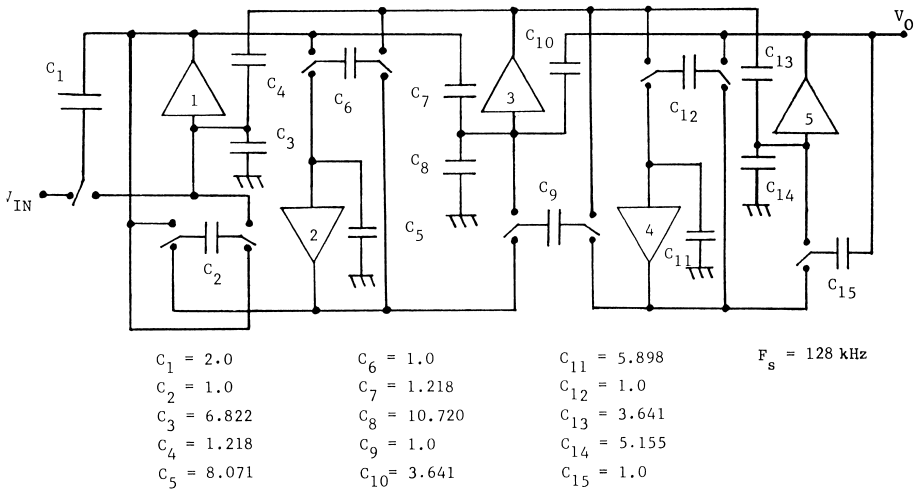


FIGURE 23 A fifth-order elliptic low-pass ladder realization based on unit-gain buffers.

while Fig. 23 shows a third design based on an approach using all unit gain buffers, rather than op amps. As can be seen from the frequency analyses in Fig. 24, all of these designs are capable of approximating the desired response very accurately (the Q enhancement visible in the state variable design can be corrected by adjusting the values of the terminating switched capacitors if desired). To compare these structures further, sensitivities were calculated for every capacitor, op amp gain, and unity buffer gain. The elements were classified as not very sensitive (0.0 to ±0.3), moderately sensitive (±0.3 to ±0.6), or very sensitive (±0.6 to ±1.5) over the pass-band frequencies 0.0–2.0 kHz, as tabulated in Table IX. The unity gain buffers are among the most sensitive elements. Although it is difficult to make definitive claims for a superior circuit, it appears that with regard to sensitivities the original state-variable ladder is the preferred circuit because it contains no element in the most sensitive category.

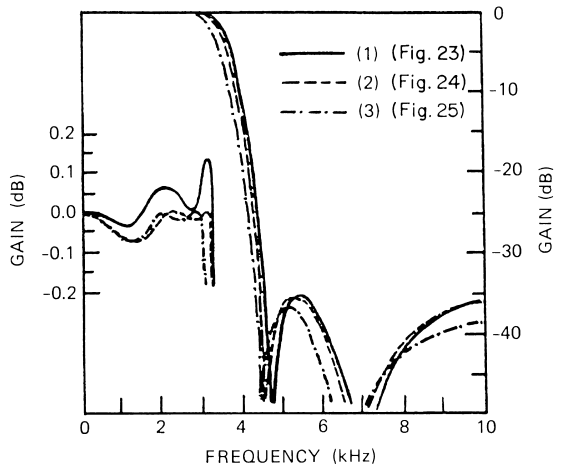


FIGURE 24 Frequency response for the three low-pass ladders of Figs. 21–23.

VII. SWITCHED ANALOG FILTERS

The biggest problem with the existing monolithic integrated-circuit technology for frequency-selective filters is the wide variation of the relative resistor values in a single integrated circuit. Variations in processing parameters (temperature, impurity dose, time, etc.) can cause the resistivity of diffused resistors to vary more than 1% on the same chip. Furthermore, this variation is uncorrelated to the variation of capacitance per unit area. Together these two effects cause RC products to vary by a factor of two when using the best processing available.

TABLE IX Sensitivities of the Most Critical Components in the Ladder Realizations of Figs. 21–23

	Output magnitude sensitivity with respect to P_k		
Circuits	**0 to ±0.3**	**0 to ±0.6**	**0 to ±1.5**
State variable Ladder Filter	C2, C6, C7, C8, C9, C12, C14, C15, C16, C17, and OP AMP 1–5	C1, C3, C4, C5, C10, C11, C13, and C18	
Inductor simulation Ladder Filter	C1, C2, C3, C4, C5, C6, C7, C10, C11, and OP AMP 1–2	C8, C9, C12, C13, C14, and C15	Buffer 1–3
Unit gain Buffer Ladder Filter	C1, C2, C3, C4, C5, C6, C7, C8, C9, C10, C11, C12, C13, C14, and C15	Buffer 3 and buffer 5	Buffer 1–2 and buffer 4

The growth of silicon dioxide on silicon is subject to similar processing variations (temperature and time), but in this case impurities are not added. For this reason the thickness of thermally grown oxide layers on a wafer does not change very much across the wafer surface. In fact, capacitors in MOS technology can be matched within 0.1%. Thus any filter topology with a transfer function dependent on capacitor ratios is ideal for unipolar (MOS) IC fabrication.

Researchers have introduced techniques that eliminate resistors by replacing them with field-effect transistors (MOSFETS) biased in the linear (ohmic) region of operation. Future development of this technique, or similar techniques that eliminate diffused resistors from filter circuits, raises the possibility that integrated filters containing resistors, capacitors, op amps, and switches may offer a simple and inexpensive means of digitally tuning state-variable analog filters. At switching rates greater than 1 MHz, MOS switched capacitor filter performance begins to degrade, which makes the use of switched-capacitor filters impractical above the audio range (20 kHz). Also, the noise associated with all MOS switches and MOS amplifiers makes the dynamic range too small for demanding high-fidelity audio applications. Another dynamic range limitation is the fundamental low breakdown voltage of MOS transistors. Current technologies do not allow signals above approximately 8 V peak-to-peak. For these reasons, this more general class of switched resistor–capacitor filters may become more practical in the future.

Figure 25 shows the prototype of a switched resistor–capacitor biquad that has been studied extensively at the University of Illinois. It is essentially a state-variable biquad active filter structure, with one analog switch added to control the bandwidth (S_b) and two analog switches added to control the center frequency (S_f). Note that the filter has both a bandpass output (V_{BP}) and a low-pass output (V_{LP}). One of the important features of this circuit is that the bandwidth and center frequency are tuned by the duty cycles on the switches, not by the switching frequencies. This is important in practical integrated circuits because it is a much simpler matter to adjust duty cycles by proper gating, whereas it is a more costly matter to change the clock frequencies over a continuous range. This prototype was built from discrete components and has not, to the author's knowledge, ever been implemented in integrated form.

Figures 26–28 show experimental results that illustrate the electronic tuning capability of this circuit. Figure 26 shows 16 different frequency responses superimposed on the same axis: A, B, C, and D are four different center frequencies, and the four curves around each of these center points represent filters with four different bandwidths. There is a slight asymmetry visible in these curves due to nonideal effects of the discrete components used in the construction. Figure 27 shows tuning for three different center frequencies with the Q of the filter held constant (i.e., both S_b and S_c at 25, 50, and 75% duty), and Fig. 28 shows tuning for three different bandwidths (S_b at 25, 50, and 75% duty) with the center frequency held constant. Experiments showed that a very wide range of center frequencies and bandwidths could be obtained by the proper choice of the switch duty cycles.

This example illustrates the tremendous flexibility that can be achieved with analog filters that contain resistors, capacitors, op amps, and switches. The combination of electronically tunable analog circuits with digital computer controls results in flexible and cost-effective

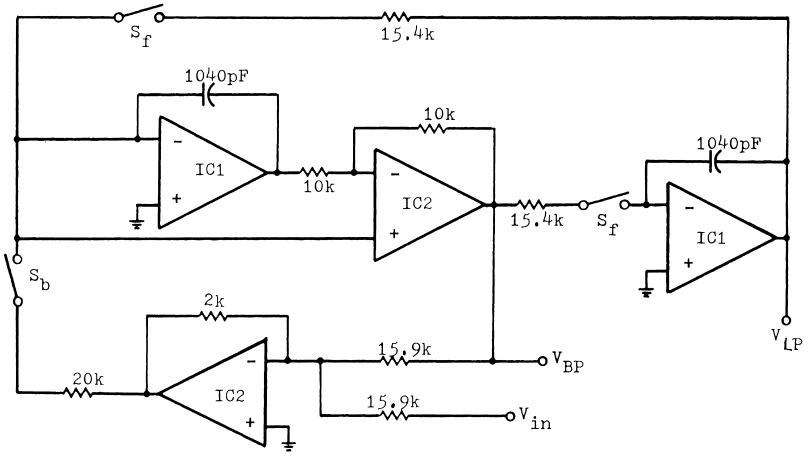


FIGURE 25 Experimental prototype for a tunable switched resistor–capacitor filter. IC1 = IC2 = LF353. All switches are part of CD4016 CMOS switch. All resistor values are in ohms.

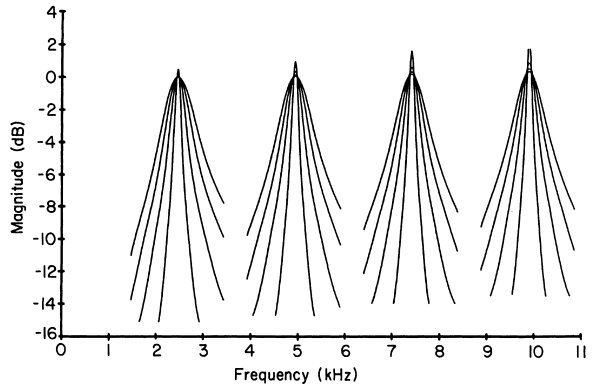


FIGURE 26 Demonstration of electronic tunability of the center frequency and bandwidth for the filter in Fig. 25.

solutions to practical signal processing problems faced by designers of electronic instrumentation.

VIII. OPTICAL SIGNAL PROCESSING

Radar, or radio detecting and ranging, has been used for decades to detect distant objects, to measure their velocity, and for terrain mapping. The basic theory first appeared in the early 1900s and was demonstrated in practice during the mid-1930s. Since its early use in World War II to detect airplane bombers, it has been refined and used in hundreds of applications, such as space exploration, aircraft avoidance systems, insect and bird tracking, weather observation and prediction, and, of course, highway speed enforcement.

When radar is used to simply detect the range to a distant object, a series of electromagnetic pulses is transmitted in the direction of the object and the two-way time delay of the reflected pulses is measured. Because it is known that the pulses propagate at the speed of light, the range to the object can easily be calculated from this

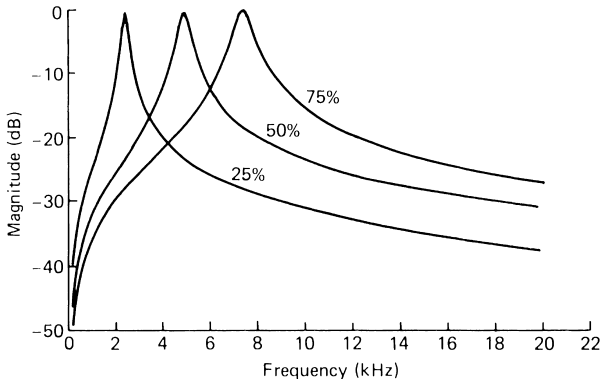


FIGURE 27 Experimental constant-Q bandpass frequency response for the filter of Fig. 25 (% indicates the switch duty cycle).

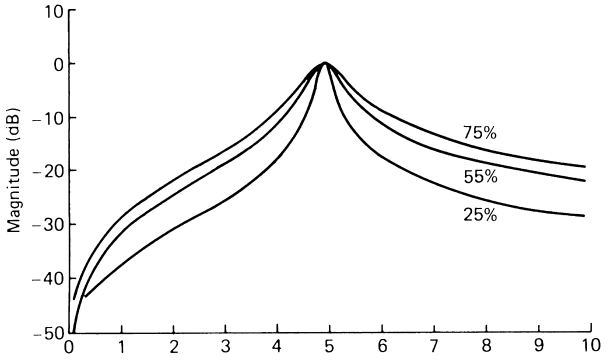


FIGURE 28 Experimental bandwidth tuning of bandpass frequency response (% indicates the switch duty cycle).

delay. If it is desired to measure the radial velocity of an object, a continuous-wave (cw) electromagnetic signal (a simple sinusoid) is transmitted and the frequency shift of the returned signal, induced by the moving object, is measured. It is a straightforward procedure to calculate the velocity of the object from the Doppler shift and the known propagation velocity of the signal.

When the azimuthal position (perpendicular to range) is also desired, the transmitted signal must be sent as a narrow beam that is no wider than the desired azimuthal resolution. It is easily shown that at a range R and carrier ω_c, an antenna with an aperture size D will have an azimuthal resolution on the order of $R\omega_c/D$ m. For a radar operating in the conventional microwave region (10^9 Hz), this implies that very small resolution requirements dictate a physical antenna of very large size, that is, several hundred feet across. For an airborne antenna, this is impractical; however, very high azimuthal resolution can be achieved by synthesizing a large antenna through recording and signal processing.

Synthetic aperture radar (SAR) is a means of obtaining high-resolution terrain maps (complex reflectivity maps) by coherently processing the backscattered radar signal's phase (referred to sometimes as a phase history). In 1953, the University of Illinois experimentally demonstrated strip-mapped SAR. This was the earliest form of SAR in which the airborne radar antenna is held at a fixed squint angle relative to the terrain patch of interest. Coherent processing of the returned radar signal produces an effective aperture many times the size of the physical antenna, thus creating an azimuthal beam width proportionally narrower, resulting in finer azimuthal resolution.

A higher resolution form of SAR is called spotlight-mode SAR. This mode requires that the antenna continually be steered toward the center of the terrain patch during the aircraft flight. The same ground area is effectively pictured from many different angles and when coherently processed, produces an image of much higher resolution

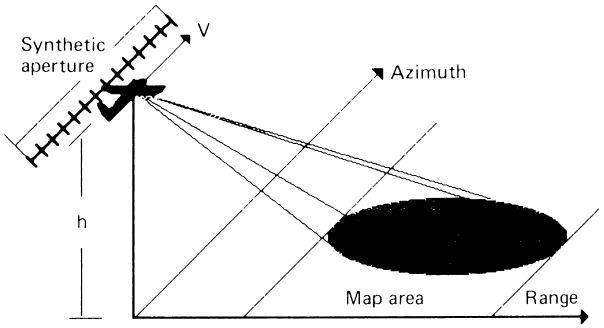


FIGURE 29 Physical geometry of a strip-mapping SAR system.

than that which could be obtained from a single observation angle. Both strip-mapping and spotlight-mode SAR can also operate in space with a low-altitude satellite orbit; however, in this case the curved shape of the orbital trajectory becomes a significant factor and must be compensated as part of the signal processing required to produce the final image.

Synthetic aperture radar image processing can be done by either optical or digital signal processing. The following discussion describes the fundamentals of optical SAR processing, a subject that represents a successful application of two-dimensional analog signal processing. Figure 29 depicts a radar platform traveling with velocity V at altitude h and carrying a side-looking antenna that illuminates an area on the ground to be imaged (map area). It is assumed in this discussion that the radar is operating in the strip-mapping mode, so the antenna is held fixed on the aircraft and the ground strip is mapped out by the forward motion of the aircraft. The radar transmits a linear FM pulse

$$s(t) = \begin{cases} \cos(\omega_c t + \nu t^2/2), & |t| \le T/2 \\ 0, & t \text{ otherwise} \end{cases} \tag{16}$$

at N distinct points along the flight path. At each location the radar receives the echo from the Map Area and records this return for later processing. Ideally, the returned signal is demodulated into a base-band (complex) representation and then recorded for later processing. If digital processing is to be used, the I and Q channels of the base-band representation are sampled and the complex samples are stored in a digital memory. If optical processing is used, the demodulation does not translate the signal completely into the base-band, but rather the signal is left on a small offset frequency so it can be represented with real analog levels and recorded on film. This is done by using the demodulated result to modulate a cathode ray tube (CRT) sweep, which writes the signal across a recording film. After the data recording is developed it becomes the input data into an optical correlator, which match-filters (compresses) the signal in both the azimuth and range dimensions to produce a high-resolution (fully focused) image.

The essential components of the coherent optical processor used to process strip-mapping SAR images are shown in Fig. 30. The film on the left is the raw recording, which contains range sweeps written side by side from top to bottom (vertically) in the figure. The cylindrical lens removes the quadratic phase modulation in the range dimension, which was placed in the transmitted waveform by the transmitter. The conical lens performs the range-dependent quadratic-phase correction in the azimuth dimension so that all range cells are brought simultaneously into focus. The spherical lense produces a Fourier transform in the back focal plane, where the second recording film is exposed through a vertical slit as it scrolls in synchronism with the raw-data film. The combination of the cylindrical and spherical lenses implements the quadratic phase-matched filter in range, while the conical and spherical lenses implement the range-dependent quadratic phase correction and integration in azimuth. The vertical slit is required to pass only the dc component of the spectrum, which means that the spherical lens performs a simple integration in azimuth when its output is viewed through the slit. Figure 31 shows the final configuration for an optical system to process SAR images.

One of the attractive features of optical signal processing for SAR is the extremely high processing rates that can

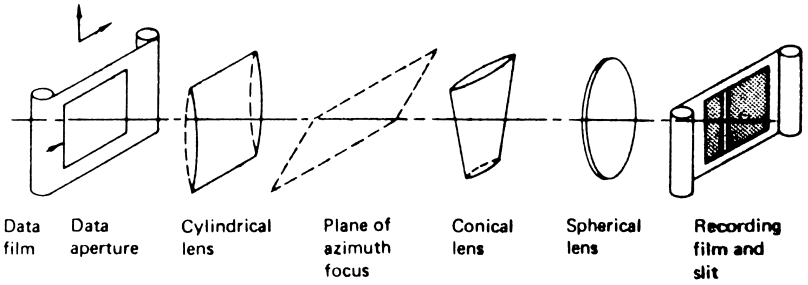


FIGURE 30 Elements of a coherent optical system for processing strip-mapping SAR imagery. [From Harger, R. O. (1970). "Synthetic Aperture Radar Systems: Theory and Design. Academic Press, New York.]

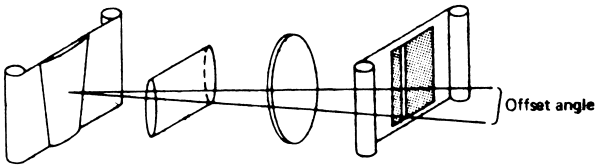


FIGURE 31 Consolidated optical system for SAR processing. [From Harger, R. O. (1970). "Synthetic Aperture Radar Systems: Theory and Design." Academic Press, New York.]

be achieved with very wide band data. Equivalent digital systems require large memories and very high speed digital processors to accomplish the same task. Unfortunately, the fact that data films must be exposed and developed, together with the fact that the optical processing equipment is not very portable, has resulted in more attention being devoted to digital processing for SAR.

SEE ALSO THE FOLLOWING ARTICLES

ANALOG ELECTRICAL FILTERS • ANALOG-SIGNAL ELECTRONIC CIRCUITS • DIFFERENTIAL EQUATIONS, ORDINARY • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL • SYNTHETIC APERTURE RADAR • ULTRA LARGE SCALE INTEGRATION (ULSI) DESIGN

BIBLIOGRAPHY

Broderson, R. W., Gray, P. R., Hodges, D. A., Allstott, D., and Jacobs, G. MOS switched capacitor filters. *Proc. IEEE* **67,** 61–75.

Embree, P. M. (1983). The analysis and performance of a tunable state variable filter employing switched resistor elements. *Proc. 1983 Int. Symp. Circuits and Systems, Newport Beach, California, May 1983*, pp. 857–860.

Fang, S. F., Tsividis, Y. P., and Wing, O. (1983). SWITCAP: A switched-capacitor network analysis program—Part I: Basic features. *Circuits Systems Mag*. **5**(3), 4–10. Part II: Advanced applications, *Circuits Systems Mag*. **5**(4), 41–46.

Fettweis, A. (1981). Switched-capacitor filters: From early ideas to present possibilities. *Proc. 1981 Int. Symp. Circuits and Systems, Chicago, April 1981*, pp. 414–417.

Fried, D. L. (1972). Analog sample-data filters. *IEEE J. Solid-State Circuits* **SC-7**, 302–304.

Gabel, R. A., and Roberts, R. A. (1980). "Signals and Linear Systems," 2nd ed. Wiley, New York.

Harger, R. O. (1970). "Synthetic Aperture Radar Systems: Theory and Design," Academic Press, New York.

Jacobs, G. M., Allstot, D. J., Broderson, R. W., and Gray, P. R. (1978). Design techniques for MOS switched capacitor ladder filters. *IEEE Trans. Circuits Systems* **CAS-25,** 1014–1021.

Jenkins, W. K. (1983). Observations on the evolution of switched capacitor[circuits. *Circuits Systems Mag*. **5**(4), 22–33.

Laker, K. R., and Fleischer, P. E. (1979). A family of active switched capacitor biquad building blocks. *Bell System Tech. J.* **58,** 2235–2269.

Vandewalle, J. (1981). Survey of computer programs for CAD of switched capacitor circuits. *Proc. 1981 Eur. Conf. Circuit Theory and Design, The Hague, August 1981*, pp. 1071–1072.

Van Valkenburg, M. E. (1974). "Network Analysis," 3rd ed. Prentice-Hall, Englewood Cliffs, New Jersey.



Signal Processing, Digital

Fred J. Taylor
University of Florida

I. Origins of Signal Processing
II. Origins of Digital Signal Processing
III. Signal and System Representation
IV. Digital Transforms
V. The Fast Fourier Transform
VI. Other Transforms
VII. Discrete Systems
VIII. Digital Filter Design
IX. Implementation
X. Hardware
XI. Applications

GLOSSARY

Aliasing Two or more distinctly different signals having identical or impersonating sample values.

Convolution Mathematical process, appearing in linear or circular form, that models the input–output filtering process.

Digital filter Digital device that is capable of altering the magnitude, frequency, or phase response of a digitally encoded input signal.

Digital signal processing chip Electronic chip or chip set designed to perform the digital signal processing operations of transforming and filtering.

Discrete Fourier transform Mathematical method of transforming a time series into a set of harmonics in the frequency domain and vice versa.

Fast Fourier transform Popular algorithm used to implement discrete Fourier transforms.

Finite word length effects Filter or transform behavior that can be attributed to the rounding or truncation of data found in finite-length register machines.

Impulse response Response of a filter to a single input sample. If of finite duration, it is said to be produced by a finite impulse response filter, otherwise an infinite impulse response filter.

Limit cycling Filter behavior that results in unwanted oscillations due to data rounding, truncation, or saturation.

Nyquist sample rate Sample rate above which a band-limited signal can be reconstructed from its sample values.

Quantization Process by which a real variable is converted into a digital word of finite precision.

Sample value Number or value associated with a single distinct sample; it is real for a discrete system and of finite precision in a digital signal processing system.

Time series Continuous string of sample values.
Window Process of isolating and modifying data to the exclusion of all others.
***z*-Transform** Mathematical method of analyzing and representing discrete signals.

DIGITAL SIGNAL PROCESSING (DSP) is a relatively new technical field that is concerned with the study of systems and signals with respect to the constraints and attributes imposed on them by digital computing machinery. It differs from other signal processing technologies—namely, analog and discrete—in terms of signal definition, computational procedures, and performance limitations. Because of the organization of a digital computer and the finiteness of digital calculations, the field of DSP has developed a unique set of analysis and synthesis procedures. A DSP system will accept a string of digitally encoded samples, called a digital time series (or time series), and modify, manipulate, classify, or quantify them using DSP algorithms. The tools used to implement these DSP algorithms are digital computer software, hardware, or their mix. The digitally processed data can then be used or analyzed by human or machines. Digital signal processing has become an essential element in the study of communication, control, medicine, speech, vision, radar, sonar systems, plus a host of other applications where the high speed and high precision of digital computers can be applied to signal and system synthesis or analysis.

I. ORIGINS OF SIGNAL PROCESSING

The foundations of DSP are interwoven with the origins of its counterparts, analog signal processing and algorithms. An algorithm is a computing procedure that is generally well suited for execution on a digital computer. The theory of algorithms can be traced back to the seventeenth- and eighteenth-century work of the celebrated mathematicians Sir Isaac Newton and Karl Freidrich Gauss. An important signal processing tool now called the discrete Fourier transform (DFT) may have been derived by Gauss as an algorithm as early as 1805. This would predate Fourier's discovery in 1807 of the infinite harmonic series, which was published in 1822. In fact, many early digital signal processing successes were simply classic numerical analysis algorithms programmed for execution on a general-purpose digital computer, but modified to exhibit special signal processing attributes.

The other foundation on which digital signal processing is built is its predecessor, analog signal processing. An analog signal would have a graph that is continuously defined along both the independent (e.g., time) and dependent (e.g., amplitude) axes. This science goes back to the very early days of electronics, when now-primitive vacuum tube amplifiers, separately packaged capacitors, inductors, resistors, and the like were used to filter signals on the basis of their frequency content. These analog systems were often modeled in terms of ordinary linear differential equations. During the 1940s it became popular to study such systems with a differential equation analysis tool called the Laplace Transform. Using these classical analog methods, a number of important filter forms were developed. Those that have had a major impact on the design of digital filters have been the maximally flat (Butterworth) filter and the equal-ripple filters referred to as Chebyshev and elliptic filters. In the late 1940s and 1950s a new dimension in analog signal processing evolved, which was called sampled data. A sampled data signal would have a graph whose dependent axis is continuously defined but whose independent axis is discrete. This is the result of sampling a continuous signal, say $x(t)$, at periodic intervals of time and then saving the resulting sampled values. For example, if a ± 10-V signal is sampled at the sample rate of 1000 times per second, the resulting discrete signal could look like $x(n) = \{\ldots, +1.23334, -0.342344, +7.24324, \ldots\}$, where the three displayed *real* sample values represent the actual value of $x(t)$ at times $t = (k-1)t_s$, kt_s, $(k+1)t_s$ where $t_s = 1/1000 = 10^{-3}$ see. These early discrete systems were used in low-frequency control applications, such as autopilots, where conventional analog filters were too bulky, unreliable, and expensive to be considered viable design options. Mathematically, the designers of discrete systems used essentially those techniques developed for linear continuous system analysis by modifying the Laplace transform to a form we now call the z-transform.

During the late 1940s another major milestone was achieved. Claude Shannon, of Bell Laboratories, and his associates noticed that if a signal $s(t)$ contains no frequency components at or above f_N Hz [$s(t)$ is then said to be bandlimited to f_N Hz], then $s(t)$ can be completely reconstructed (interpolated) from its sample values provided the samples are taken at a rate equal to or in excess of $f_s = 2f_N$ samples per second. This condition is known as the Nyquist sampling the orem, f_s is referred to as the Nyquist sample frequency, and f_N is sometimes called the Nyquist frequency. In the time domain, the samples must be spaced no further apart than $t_s = 1/f_s$ sec, where t_s is the sample period. If a bandlimited signal is sampled at a rate slower than the Nyquist sample frequency, a type of error called aliasing can occur. As the name suggests, an aliasing signal impersonates another signal in such a manner that they cannot be discriminated from each other. This problem is suggested in Fig. 1. For example, if two signals $s_1(t) = 1$ and $s_2(t) = \cos(2\pi t)$ are sampled once

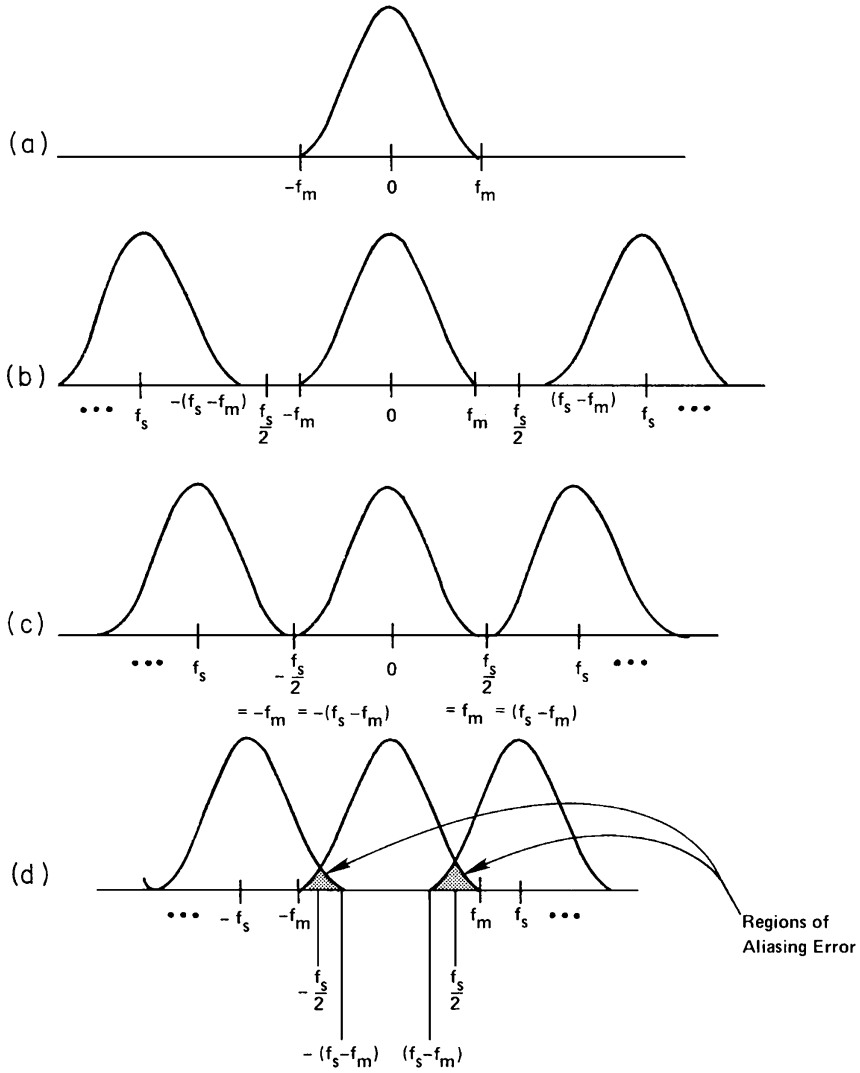


FIGURE 1 Example of aliasing: graphical interpretation of spectral isolation and overlap as a function of sample rate. (a) Signal spectra. (b) Spectra for sampling beyond the Nyquist sampling rate; $f_s > 2f_m$ (f_m is the maximum frequency). (c) Spectra for sampling at the Nyquist rate ($f_s = 2f_m$). (d) Spectra for sampling below the Nyquist rate ($f_s < 2f_m$).

per second, their resulting sample series would both be given by {1, 1, 1, . . .}. The reason for this undesirable circumstance is that $s_2(t)$ is bandlimited to 1 Hz and therefore must be sampled at a rate in excess of two samples per second. Once properly applied, the sampling theorem had an immediate impact on the fields of communication, control, as it later did on the new field of digital signal processing.

II. ORIGINS OF DIGITAL SIGNAL PROCESSING

The study of computing algorithms predates the advent of the digital computer. However, once digital computing machines became available, there was immediate interest in implementing a host of computing algorithms (some dating from the seventeenth and eighteenth centuries). These early attempts were, in a large part, extensions to the then maturing field of numerical analysis. Because the digital computer was arithmetically fast and could be programmed to iteratively solve a complex problem, computing algorithms were extensively used to statistically analyze and mathematically manipulate signals. During the 1960s the race into space accelerated this type of signal processing and gave rise to a host of estimation, smoothing, and prediction algorithms for navigation, control, and space communication applications. The fields of pattern recognition and statistical hypothesis testing (e.g., radar

signal detection) were also greatly enriched with the availability of general-purpose computers.

During the transition period, ranging through much of the 1960s, the general-purpose digital computer was too expensive and unreliable to be used as an on-line signal processing tool. But that began to change. In the mid-1960s the first DSP milestones were achieved in the form of fast transforms and filters. One of the attempts to use the computer as a signal processing tool was in the area of spectral analysis using Fourier transform and its variations. A discrete Fourier transform accepts as input a set of discrete samples in the time domain (i.e., a time series) and transforms them into information in the frequency domain. Unfortunately, because of its high multiplication budget, the DFT formula was not amenable to efficient digital computer data processing. It was Cooley and Tuckey who popularized the so-called fast Fourier transform (FFT), which revolutionized the field of the DFT. The FFT achieved two milestones. First, it provided a service—namely, faster spectral analysis. Second, it was designed with an awareness of the strengths and limitations of digital computers—namely, memory utilization and data flow. Because of this, the FFT altered the attitude of digital signal processing engineers and scientists, who charted a new direction thereafter. Refinements and design procedures that optimize computer execution rates and resource utilization were sought. Digital computer architectures have also been developed that reflect the requirement of DSP algorithms. Finally, many other fast digital algorithms and transforms have since been discovered.

The other major branch of DSP theory has concerned digital filters, which can replace many of the analog electronic filters that have been in daily use for over half a century. Passive analog filters are designed with resistors (R), capacitors (C), and inductors (L) only. Active analog filters add electronic amplifiers to the R, L, C parts list. However, these classic filters must be designed with realistic values for resistance, capacitance, and inductance. Because of the physical limitations of passive components, analog filters have a limited frequency response. They also perform poorly at very low or very high frequencies. Analog filters often require extensive and costly alignment and adjustment and require maintenance throughout their useful life span. Linear phase behavior is difficult to achieve, and the fact that analog filters cannot be programmed limits their utility and design flexibility. The alternative digital filter technology overcomes these limitations. Digital filters are a by-product of the great advances in solidstate electronics during the 1970s, which produced a wealth of high-performance, low-cost electronic devices capable of manipulating signals in an algorithmic manner not possible with analog units. Advances in digital technology such as the microprocessor have created the new field known as digital signal processing.

From a theoretical viewpoint, digital filters provide a rich field for study. Because of their robust theoretical base, one would expect more advanced digital signal processing systems to be developed in the future that are based on mathematical abstractions. The level of mathematical consciousness in digital processing is attributable to the nature of the hardware in which these systems reside. Whereas analog systems are defined over a *real coefficient field*, the data and coefficients associated with a digital filter are defined with finite precision and stored in registers of finite word length. The levels of mathematical skill required to work algebraically in this area are more advanced than those required in the field of real numbers.

The advantages of digital filters are numerous:

1. They can be fabricated with available high-density, low-cost digital hardware. This should continue to improve as digital signal processing assimilates the continuing advance in the digital electronics industry.
2. Certain classes of digital filters have guaranteed stability.
3. They are free of the impedance-matching problems associated with analog filters.
4. Digital filters can work at extremely low frequencies that cannot be supported by analog filters.
5. Digital filters are programmable if realized with a programmable processor.
6. They efficiently support computer-aided data analysis [simple input–output (I/O) logging and transfer] and can often be interfaced to the power supplies and mechanical structure of existing digital processors.
7. They can work over a wide range of critical frequencies, which is a difficult task for analog filters.
8. They can be used in conjunction with data compression (i.e., input and/or output) schemes.
9. Certain digital filters have outstanding phase linearity.
10. They can have high accuracy and precision. The precision of an analog filter generally does not exceed 60–70 decibels. The precision of a digital filter can be extended simply by increasing word length. For example, an n-bit digital word defines a dynamic range approximated by $6.0206n$ dB.
11. Digital filters do not require periodic alignment, which is necessary for analog filters. Furthermore, they do not "drift" with parameter aging or environmental changes.
12. Because of their digital nature, digital filters are less sensitive to certain classes of noise that corrupt analog filters (e.g., line frequency noise).

The disadvantages of digital filters are few:

1. They are subject to quantization errors.
2. They may exhibit some form of limit cycling.
3. Hardware development time may be longer.
4. Their design and synthesis may be more difficult algebraically unless a general-purpose digital computer is used to support the design task.

Many of these disadvantages can be overcome through good digital design practices and familiarity with hardware, analysis protocols, and procedures.

General-purpose computers became available to scientists during the 1950s. At first, primitive filters were designed and implemented as computer programs. Some simply emulated the calculus operation of integration by use of iterative data processing techniques. An iterative calculation is one in which the output is a function of both the input and past outputs. Because of the recurrence relationship, these algorithms would now be called recursive filters. If the output can be written as a function of input or independent variables only, a filter would be called nonrecursive or, sometimes, a transversal filter. Digital filters can be used to smooth, interpolate, condition, or analyze data. For example, the simple recursive filter given by $y_{n+1} = -0.9y_n + x_n$ amplifies signals of alternating sign and attenuates signals of constant sign. Starting at $y_0 = 0$, the signal $x_n = (-1)^n$ produces $y_1 = 1$, $y_2 = -1.9$, $y_3 = 2.71$, and so forth, while $x_n = 1$ yields $y_1 = 1$, $y_2 = 0.1$, $y_3 = 0.9, \ldots$ Scientific data analysis tools of this type operate "off-line" and use floating-point arithmetic. However, they do not represent true DSP statements since they were not designed with an awareness of the imprecision of digital arithmetic, memory limitations, computational complexity, data and control flow, and the like. But it was not long before engineers and scientists were developing computer routines and algorithms that did lend themselves to fast and efficient digital processing. This began the era of digital signal processing.

III. SIGNAL AND SYSTEM REPRESENTATION

Signal processing is concerned with the problem of manipulating and managing information. The information source (signal) may appear in one of several forms. A continuous signal process is one that is continuously resolved in both its independent and dependent axes. A discrete signal process is discretely resolved (quantized) only in the dependent variable. A digital signal process is discretely resolved in both variables. Periodic (uniform) sampling is the most common method of converting an analog signal to a discrete format.

Shannon's theorem (*ca.* 1948) states that if $x(t)$ is a signal whose highest-frequency component is bounded by f_N, and if $x(t)$ is periodically sampled sot that the sample period t_s is less than or equal to $t_s < 1/2f_N$ (referred to as the Nyquist sample rate), then $x(t)$ can be interpolated from its time series by the rule

$$x(t) = \sum_{n=-\infty}^{\infty} \frac{x(nt_s)\sin[\omega_c(t - nt_s)]}{\omega_c(t - nt_s)}.$$

However, use of this interpolating equation is not practical, and the following simpler approximations to the equation have been developed:

The zeroth-order hold (interpolator):

$$x(t) = x(n), \qquad t \in [nt_s,\ (n+1)t_s].$$

The first-order hold (interpolator):

$$x(t) = x(n) + t[x(n+1) - x(n)]/t_s,$$
$$t \in [nt_s,\ (n+1)t_s].$$

The zeroth-order hold is often called a sample and hold circuit, while the first-order is called a linear interpolation.

The analysis of a discrete system can be performed using transform methods. One of the popular discrete transforms is the z-transform. It belongs to a class of algebraic operations referred to as impulse-invariant transforms, which means that the inverse transform gives back the original sample values. The z-transform is related to the Laplace transform variable s by $z = \exp(st_s)$. If $\{x(n)\}$ is a time series consisting of sample values $x(n) = x(nt_s)$, then its z-transform is given by

$$x(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

Some of the basic properties of the z-transform are as follows:

1. The z-transform of $\{x(i)\}$ exists if and only if $\{x(i)\}$ is unique and bounded for all n.
2. Two z-transforms $X(z)$ and $Y(z)$ are equal if and only if $\{x(i)\} \equiv \{y(i)\}$.
3. The z-transform of $x(t/T)$ is independent of the sampling period T.
4. The z-transform is a linear operator in that if $x(n) \to X(z)$ and $y(n) \to Y(z)$, then if $z(n) = x(n) + y(n)$, it follows that $Z(z) = X(z) + Y(z)$.
5. The shifting property holds in that if $y(n) = x(n+m)$ and $x(n) \to X(z)$, then $Y(z) = z^m X(z)$.
6. The initial and final value theorem states that if $x(0) = \lim X(z)$ as $z \to \infty$, $x(\infty) = \lim[1 - z^{-1})X(z)]$ as $z \to 1$.

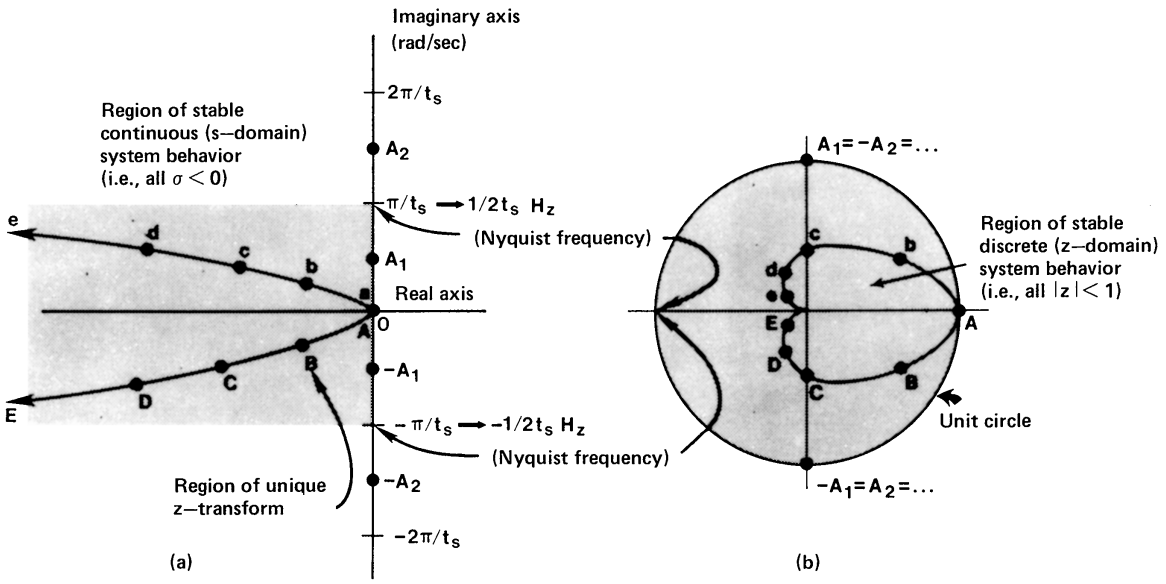


FIGURE 2 Relationship between *s*- and *z*-transforms in the complex plane. (a) Laplace or *s* domain, $s = \sigma + i\omega$; (b) *z*-transform domain, $z = e^{st}$. The shaded *s* plane in (a) is the region that results in a unique and stable *z*-transform. The shaded area of the *z* plane in (b) is the region of the stable *z*-transforms.

7. The linear convolution property holds in that if $x(n) \rightarrow X(z)$ and $y(n) \rightarrow Y(z)$, then $z(n) = x(n) * y(n)$ implies that $Z(z) = X(z)Y(z)$. A more formal mathematical statement is

$$z(n) = \sum_{k=-\infty}^{\infty} x(k)y(n-k)$$
$$= \sum_{k=-\infty}^{\infty} x(n-k)y(k) \quad \overset{Z}{\longleftrightarrow} X(z)Y(z).$$

8. Parseval's theorem becomes

$$\sum_{k=-\infty}^{\infty} x(n)y(n) = \frac{1}{2\pi j}\int X(z)Y(z^{-1})z^{-1}\,dz.$$

9. Stability properties in the *s* domain transfer into the *z* domain. The locations of stable *s*-domain pole values are given by $s = \sigma + j\omega$ for $\sigma < 0$. Substituting this condition into the defining equation for a *z*-transform, it follows that $|z| = |\exp(\sigma t_s) \times \exp(j\omega t_s)| \leq |\exp(\sigma t_s)| < 1$ if $\sigma < 0$. That is, the stable complex values of *z* are bounded to be within the periphery of the unit circle given by $|z| = 1$. This condition is often referred to as the circle criterion and is summarized in Fig. 2.

Properties 7 and 8 make the *z*-transform well suited for representing systems as well as signals.

There are some restrictions on the use of the *z*-transform. From the definition of *z* [i.e., $z = \exp(st_s)$] it is seen that the *z*-transform is unique if and only if the modulus of st_s is bounded by $\pm\pi$ (i.e., $|st_s| \leq \pi$). For $s = j\omega = j2\pi f$, it follows that $|2\pi f t_s| \leq \pi$ or $|f| \leq 1/2t_s$, which is the Nyquist sampling frequency. It is sometimes useful to interpret the *z*-transform terms of parameters such as those shown in Table I.

Finally, the inversion of a given $X(z)$ is formally defined in terms of the contour integral

$$x(n) = Z^{-1}[X(z)] = \frac{1}{2\pi j}\oint_C \frac{X(z)z^n\,dz}{z},$$

where *C* is a restricted closed path found in the *z* plane. However, one rarely approaches the problem of inversion through the use of this difficult integral equation. Instead, more simplified methods have been developed that expedite the inversion process. The most popular of these methods are

TABLE I Parameters Used in Interpreting the *Z*-Transform

Parameter	*z* Operator	Range
f, Hz	$z = e^{j2\pi f t_s}$	$0 \leq f \leq f_s$
ω, rad/sec	$z = e^{j\omega t_s}$	$0 \leq \omega \leq 2\pi f_s$
$f' = f/f_s$ normalized to f_s	$z = e^{j2\pi f'}$	$0 \leq f' \leq 1$
$f^* = f'/0.5$ normalized to f Nyquist	$z = e^{jf^*}$	$0 \leq f^* \leq 2$
θ, rad	$z = e^{j\theta}$	$0 \leq \theta \leq 2\pi$

1. Long division.
2. Partial fraction or Heaviside expansion.
3. Residue theorem of complex variables.

The standard z-transform is by no means the only mapping of the s domain to the z domain in practical use today. Many others have been considered, and have enjoyed a certain degree of popularity such as (1) the bilinear z-transform and (2) the matched z-transform.

The *bilinear z-transform* is defined by

$$s = \frac{2(z-1)}{t_s(z+1)}.$$

A strength of this transform is its ability to map a given continuous signal having a Laplace representation $H(s)$ into the z domain without the algebraically complex operations associated with the standard z-transform. However, the bilinear z-transform is not impulse invariant and, as a result, is not necessarily the transform of choice for time domain analysis or synthesis. However, the frequency response modeling ability of a bilinear transform system is superior to that of the standard z transformed system. Here the metric of comparison is assumed to be the similarity of the frequency response of an analog filter $H(s)$ and its discrete model $H(z)$. The spectral properties of the bilinear transform are due to its ability to distort the frequency axis by a property known as warping. If Ω represents the frequency axis of a continuous system whose Laplace transform is $H(z)$ and ω is that for the discrete system $H(z)$, where $H(s) \rightarrow H(z)$ by using the bilinear transform, then the warping equation is

$$\Omega = (2/t_s)\tan(\omega t_s/2),$$
$$\omega = (2/t_s)\tan^{-1}(\Omega t_s/2).$$

The warping of the continuous frequency axis is shown in Fig. 3. If the design of a discrete filter is predicted on knowledge of its continuous filter model $H(s)$, then the final design specifications given along the discrete frequency axis are prewarped into the continuous frequency axis. From this, the resulting continuous filter $H(s)$ is designed.

The matched z-transform is normally used to model narrowband frequency-selective filters and is given by

$$H(s) = \frac{\Pi_{i=1}^{N}(s-a_i)}{\Pi_{i=1}^{N}(s-b_i)},$$
$$H(z) = \frac{\Pi_{i=1}^{N}\left[1-\exp(a_iT)z^{-1}\right]}{\Pi_{i=1}^{N}\left[1-\exp(b_iT)z^{-1}\right]}.$$

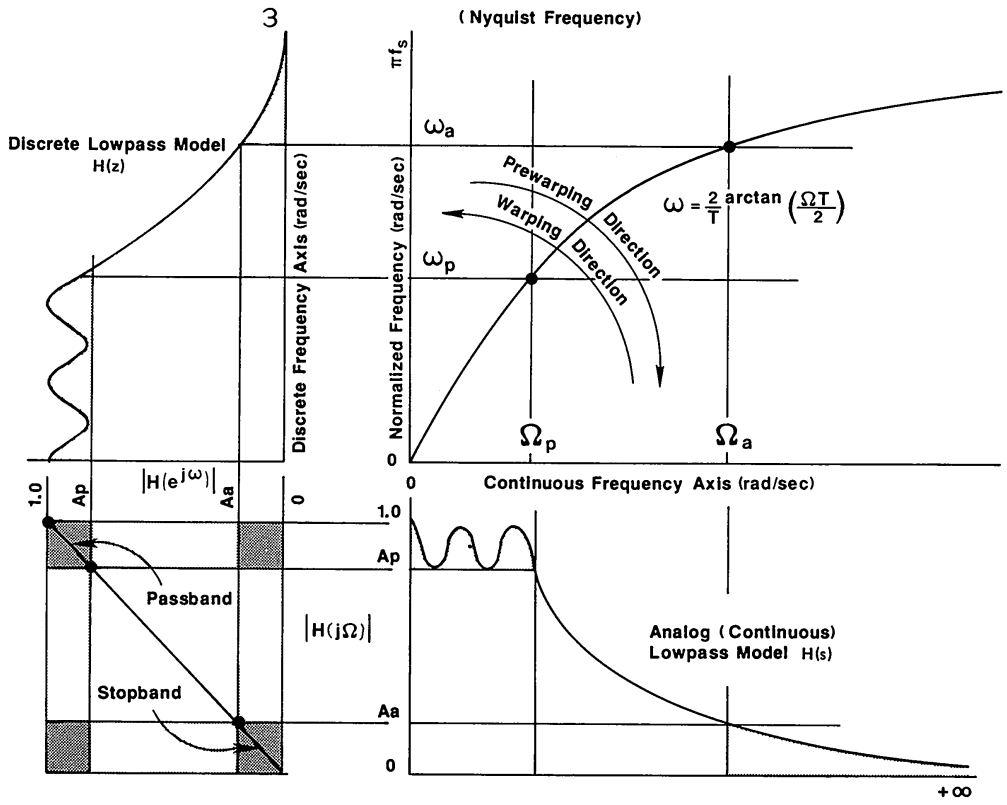


FIGURE 3 Frequency and prewarping relationship between continuous (Ω) and discrete (ω) frequency domains.

However, in most frequency domain design applications the bilinear z-transform is used, and when time domain response is the issue, the standard z-transform is preferred.

IV. DIGITAL TRANSFORMS

The analysis of a time series need not be performed only in the time domain. Other options such as the frequency domain are often used. The mapping of a signal from one domain (e.g., time) to and from another (e.g., frequency) is called a transform. Sometimes the transform has physical significance; at other times it is simply an artifact to support efficient analysis. In the field of digital signal processing, transforms are studied in the context of their problem-solving capability as well as their digital computational efficiency.

Historically, transforms have been a very rich area of study for engineers and scientists. Some, such as the Laplace transform, are used to solve complicated mathematical problems; others, like the Fourier transform, are used to establish a new analytic database in which the signal can be analyzed. The Fourier transform converts a signal into its frequency domain image called a spectrum. Analysis of the spectral signature of a signal is critical in many problems of communications, control, defense, pattern and voice processing, medicine, and other technical areas. The exponential form of the Fourier transform of a signal $x(t)$ is given by

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,dt,$$

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\,d\omega.$$

However, because of the finite storage and speed capability of a digital computer, infinitely long real or complex signals are replaced by their finite time series approximation. If the original signal is bandlimited to f_N Hz and is sampled at or above the Nyquist rate of $t_s < 1/2f_N$ samples per second for N contiguous samples, then a discrete Fourier transform can be defined and used to approximate a Fourier transform. The DFT maps an N-point time series into the frequency domain by the rule

$$X(k) = \sum_{n=0}^{N-1} x(n)W_N^{nk} \quad \text{(DFT)},$$

and the inverse DFT (IDFT) returns it to the time domain and is given by

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)W_N^{-kn} \quad \text{(IDFT)},$$

where W_N is a complex exponential $W_N = \exp(-j2\pi/N)$ and is sometimes referred to as the twiddle factor. A few of the more important DFT properties are

1. Linearity: $\text{DFT}[ax(n) + by(n)] = a\text{DFT}[x(n)] + b\text{DFT}[y(n)] = aX(k) + bY(k)$.
2. Cyclic time and frequency shifting: $\text{DFT}[x(n-r)] = X(k)W_N^{rk}$ and $\text{IDFT}[X(k-r)] = x(n)W_N^{nr}$.
3. Time reversal: $\text{DFT}[x(n)] = X(-k) = X(N-k)$. This provides a mechanism by which negative frequencies can be interpreted.
4. Symmetry for a real input: for $x(n)$ real, $X(k) = X^*(-k)$. Two special cases can be considered. If $\{x(n)\}$ is a real even time series, $X(k)$ is real (i.e., a cosine series). If $\{x(n)\}$ is odd, $X(k)$ is imaginary (i.e., a sine series).

Some of the more important DFT milestones are shown in Table II.

The DFT process of converting an N-point time series into a spectrum is indicated in Fig. 4. If $x(t)$ is a bandlimited signal whose maximum frequency component is f_M Hz, then the continuous signal must be sampled at a speed greater than the Nyquist sample rate of $2f_M$ samples/sec if aliasing errors are to be avoided. The resulting time series is then processed by the DFT formula to produce a set of N harmonics denoted by $\{X(k)\}$, where $X(k)$ is the kth harmonic and is located at frequency $f_k = k/T$ Hz. The term $T = Nt_s$ is called the fundamental period, and $t_s < 1/2f_M$ is the sample rate. The N harmonics produced by a DFT range from $f_0 = 0$ Hz to $f_{N-1} = (N-1)/Nt_s$ Hz and are discretized into N harmonics equally spaced on f_1 Hz centers. The term $f_1 = 1/T = 1/Nt_s$ is called the fundamental frequency, and $f_0 = 0$ Hz is referred to as the dc component. If $x(t)$ is a real signal, the $N/2$ most significant harmonics can be physically interpreted as negative frequency spanning the range $[-f_s/2, 0]$.

The Fourier transform is defined over an infinitely long data record spanning $t \in (-\infty, \infty)$ and is represented in Fig. 4 as $X_{FT}(f)$. In contrast, the DFT transforms a time series of finite duration defined over the time instances $t = 0, t_s, \ldots, (N-1)t_s$. Furthermore, the DFT incorporates the mathematical assumption that the transformed time series periodically repeats itself for all time with a period T. That is, if $\{x(n)\}$ is transformed into $\{X(k)\}$ the inverse transform is actually $\{x_p(n)\}$, where $x_p(n) = x(n)$ for $n = 0, 1, \ldots, N-1$; otherwise $x_p(qN+n) = x((qN+n) \bmod N)$ for q an integer. Since in general $x_p(qN+n) \neq x(qN+n)$, the spectra in question may differ. This is called the finite aperture effect. If the original signal was, in fact, periodic with period T_p, then it is possible to achieve agreement between the DFT and the Fourier transform, which is now reduced to a Fourier series. Disagreements,

TABLE II Discoveries of Efficient Methods of Computing the DFT[a]

Researchers	Date	Sequence lengths N	Number of DFT values	Application
C. F. Gauss	1805	Any composite integer	All	Interpolation of orbits of celestial bodies
F. Carlini	1828	12	—	Harmonic analysis of barometric pressure
A. Smith	1846	4, 8, 16, 32	5 or 9	Correcting deviations in compasses on ships
J. D. Everett	1860	12	5	Modeling underground temperature deviations
C. Runge	1903	$2^n k$	All	Harmonic analysis of functions
K. Stumpff	1939	$2^n k, 3^n k$	All	Harmonic analysis of functions
G. C. Danielson and C. Lanczos	1942	2^n	All	X-ray diffraction in crystals
L. H. Thomas	1948	Any integer with relatively prime factors	All	Harmonic analysis of functions
I. J. Good	1958	Any integer with relatively prime factors	All	Harmonic analysis of functions
J. W. Cooley and J. W. Tukey	1965	Any composite integer	All	Harmonic analysis of functions
S. Winograd	1976	Any integer with relatively prime factors	All	Use of complexity theory for harmonic analysis

[a] After Herdeman, Johnson, and Burris, *IEEE ASSP Magazine*, October 1984.

if any, would now be attributed to aliasing or a phenomenon called leakage, as suggested in Fig. 5. However, leakage does not occur when the signal period and sampling interval are related by $T = mT_p$, where m is an integer.

Finite aperture effects can often be minimized with the use of data windows. A data window modifies a raw time series $\{x(n)\}$ by using the memoryless rule $x_w(n) = w(n)x(n)$, where $w(n)$ is the window function. The duality theorem of Fourier transforms states that multiplication in the time domain is equivalent to convolution in the frequency domain, so it follows that $X_w(\omega) = W(\omega) * X(\omega)$. The object of a window is to place $X_w(\omega)$ in maximum agreement with $X(\omega)$, which implies that ideally $W(\omega)$ is given by the Dirac delta distribution $\delta_D(\omega)$. Since the convolution of anything with a δ_D is simply a copy of the original, it would follow that $X_w(\omega) \equiv X(\omega)$. Unfortunately, if $W(\omega) = \delta_D(\omega)$, then $w(t) \equiv 1$ for all $t \in (-\infty, \infty)$, which

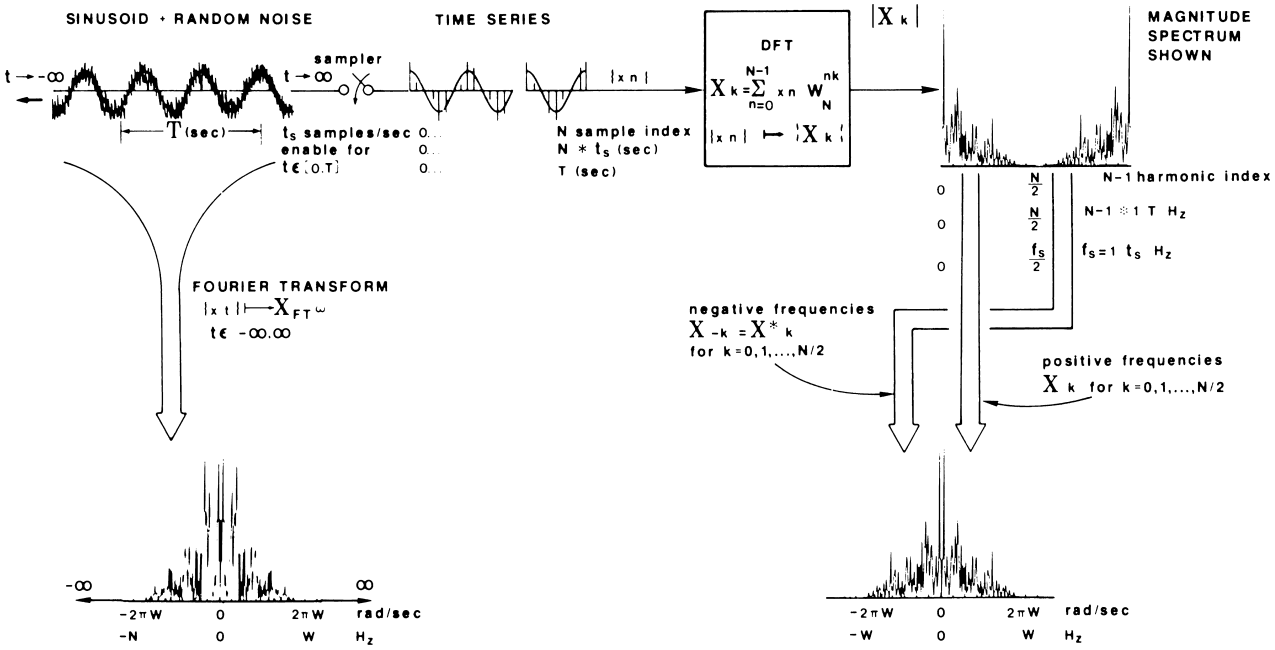


FIGURE 4 Comparison of Fourier and discrete Fourier spectra of a common signal. Note the differences between DFT and Fourier transforms due to finite aperture effects in the spectra at the bottom.

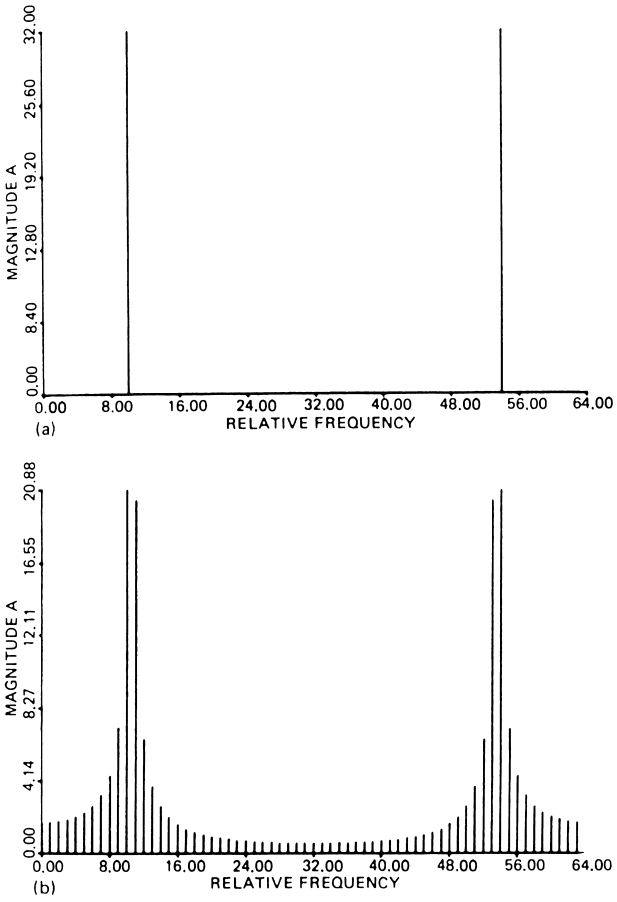


FIGURE 5 Spectra representing the DFTs of a sinusoid completing (a) 10 full cycles (no leakage) and (b) 10.5 cycles (leakage).

is unrealistic (i.e., $T \to \infty$). If T is to be of finite duration, a compromise is required. It is desired to find a finite-duration window whose spectral shape is an approximation to $\delta_D(\omega)$. Standard window functions commonly in use, which achieve this objective to various degrees, are summarized in Table III. Subjectively, the spectra produced from a windowed time series are cosmetically altered to be smoother and more interpretable than thin unwindowed counterparts. Some of these windows are shown graphically in Fig. 6.

Two-dimensional data sets are often encountered in picture and scene analysis, medicine, seismology, and so forth. A two-dimensional (2D) time series can be denoted by $x(n_1, n_2) = x(n_1 t_s, n_2 t_s)$. The sample from a 2D z-transform of the time series $\{x(n_1, n_2)\}$ is given by $X(z_1, z_2) = \sum\sum x(n_1, n_2) z_1^{-n_1} z_2^{-n_2}$, with the 2D DFT satisfying

$$X(k_1, k_2) = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x(n_1, n_2) W_{N_1}^{n_1 k_1} W_{N_2}^{n_2 k_2}.$$

By rearranging the DFT summation to read

$$X(k_1, k_2) = \sum_{n_1=0}^{N_1-1} W_{N_1}^{n_1 k_1} \sum_{n_2-0}^{N_2-1} x(n_1, n_2) W_{N_2}^{n_2 k_2}$$

$$= \sum_{n_1=0}^{N_1-1} W_{N_2}^{n_1 k_1} X'(n_1, k_2),$$

it can be seen that a 2D DFT can be computed by N_1 applications of an N_2-point DFT. The 2D IDFT would be analogously defined. Also, 2D windows are simply symmetric extensions of the one-dimensional (1D) case and can be directly applied to $x(n_1, n_2)$. The 2D DFT approximation of an ideal lowpass filter is shown in Fig. 7.

V. THE FAST FOURIER TRANSFORM

The fast Fourier transform has become a major DSP tool since being popularized by Cooley and Tuckey in 1965. In the FFT formula, the DFT equation $X(k) = \sum x(n) W_N^{nk}$ is decomposed into a number of short transforms and then recombined. The basic FFT formulas are called radix-2 or radix-4 although other radix-r forms can be found

TABLE III Commonly Used Window Types

Window type	Definition
Rectangular or uniform	$w(n) = \begin{cases} 1, & \lvert n\rvert \le (N-1)/2 \\ 0, & \text{otherwise} \end{cases}$
Hamming ($a = 0.54$), Hann ($a = 0.50$)	$w(n) = \begin{cases} 1 + (1-a)\cos(2\tau n/N), & \lvert n\rvert \le (N-1)/2 \\ 0, & \text{otherwise} \end{cases}$
Kaiser	$w(n) = \begin{cases} I_0(\beta\{1 - [2n/(N-1)]^2\}^{1/2}), & \lvert n\rvert \le (N-1)/2 \\ 0, & \text{otherwise} \end{cases}$
	where I_0 represents a zeroth-order Bessel function given by $I_0(x) = 1 + \sum_{k=1}^{\infty} [((1/k)(x/2))^k]^2$

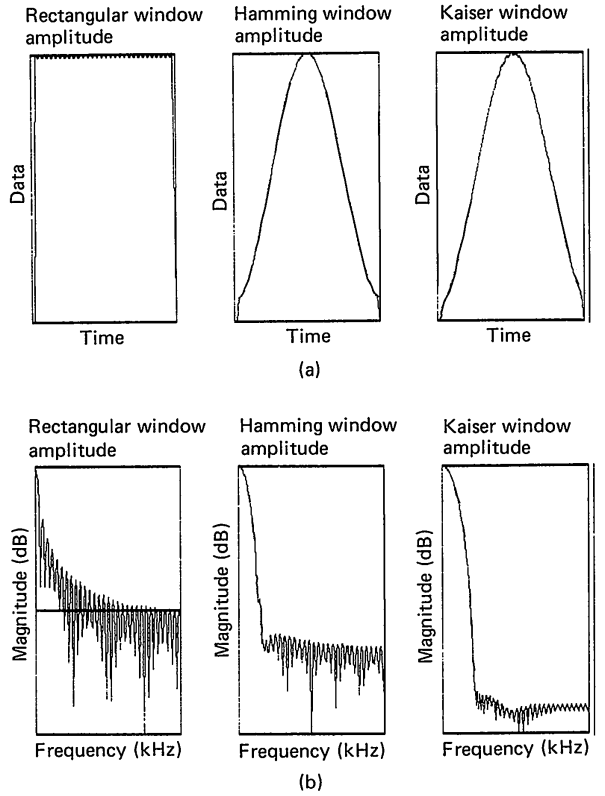


FIGURE 6 Examples of standard windows (rectangular, Hamming, and Kaiser) showing (a) time series envelope and (b) log frequency spectra. (Courtesy of the Athena Group.)

for $r = 2^k$, $r > 4$. In a radix-r implementation a problem having $N = r^n$ points is decomposed into a DFT algorithm having n/r levels, each consisting of N/r smaller r-point DFT mappings. To appreciate how the FFT algorithm works, consider the case $k, n < N = 2^n$.

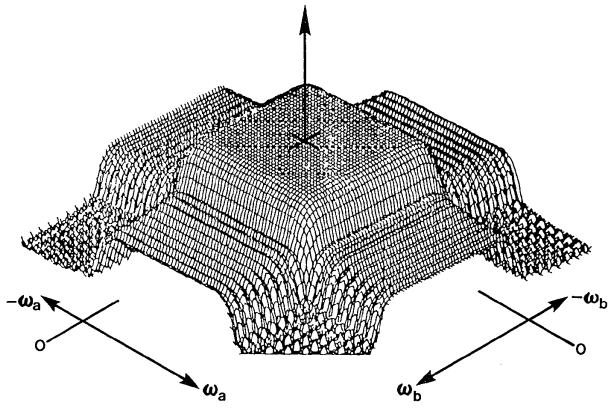


FIGURE 7 Two-dimensional DFT of a low-pass filter. The ideal low-pass filter would appear as a rectangle having infinitely steep skirts in the frequency domain.

Using the basic DFT equation, N complex multiplications are required to compute one harmonic $X(k)$. Since there are N harmonics, N^2 complex multiplications are needed. Suppose the time series $\{x(n)\}$ is decomposed into two $N/2$-point time series of even and odd samples. The complex multiplication count for each small transform is $(N/2)^2 = N^2/4$. The total complex multiplication count would then become $(N^2/4) + (N^2/4) = N^2/2$, or half the number previously required. Continuing this partitioning policy, the complex multiplication budget becomes $(N/2)\log_2(N)$, which is significantly less than the DFT direct mechanization. This reduced multiplication budget translates into speed, hence the term *fast* Fourier transform.

If the decomposition of an N-point time series takes place as an input operation, it is called decimation in time (DIT) and the DIT–FFT output spectra are sequentially ordered. If the input is left sequentially ordered, then the output is reordered in a nonsequential manner and is called decimation in frequency. For a radix-2 FFT, the decimation rule is known as bit reversing. The FFT computer code is compact and easy to implement in software. Hardware FFT realizations are ubiquitous and are generally designed for FFTs of length $N = 2^{10} = 1024$ or less. For longer transforms, block transform methods are used that synthesize a long transform by combining the outputs of a set of short transforms.

VI. OTHER TRANSFORMS

The DFT belongs to a class of transforms called orthogonal transforms, and it is not the only member of this calss used in DSP applications. Some of the more popular are the Walsh, slant, and COSINE transforms. Another class of transform, called the number theoretic transform, can be applied to fast DFTs or convolutions. A number of DFT formulas are based on optimized short transforms (e.g., Winograd DFT), which are used to grow longer transforms. The problem with these schemes is that they typically have a very complex data management structure. Finally, based on Fermat and Mersenne primes, multiplier-free transforms and convolution algorithms have been developed, but they are valid only for restricted transform lengths.

VII. DISCRETE SYSTEMS

Linear system theory is important in contemporary engineering. Linear systems may be characterized by their impulse response $h(n)$, which is the system's output to an input impulse stimulus. The output of a discrete

linear system, say $y(n)$, is given by the linear convolution equation

$$y(n) = \sum_{i=-\infty}^{\infty} h(n-i)x(i)$$
$$= \sum_{i=-\infty}^{\infty} h(n)x(n-i) = h(n) * x(n),$$

where $x(n)$ is the input. The analysis of the continuous linear convolution process can be replaced with a set of Fourier transform operations. However, an N-point DFT assumes that the transformed signal is periodic with period N samples. If $\{x(n)\}$ and $\{h(n)\}$ are two N-point time series with periodic extensions $x_p(n) = x(n \bmod N)$ and $h_p(n) = h(n \bmod N)$, then their circular convolution, denoted $x(n) \circledast y(n)$, is given by

$$x(n) \circledast y(n) = \sum_{i=0}^{N-1} x(n)y((n-i) \bmod N)$$
$$= \sum_{i=0}^{N-1} x((n-i) \bmod N)y(n).$$

In general, $\mathrm{DFT}(x(n) * y(n)) \neq \mathrm{DFT}(x(n) \circledast y(n))$. Nevertheless, there are times when one would like to use the computational power of the FFT to do linear convolution. When this is the case, a technique called zero padding can be used to create two new zero-filled $2N$-sample time series $\{x_Z(n)\}$ and $\{h_Z(n)\}$. Each consists of exact images of the original N-sample time series $\{x(n)\}$ and $\{h(n)\}$, which are appended with blocks of N zeros. Using $2N$-point DFTs and an IDFT, it is known that $\mathrm{IDFT}(X_{ZF}(k)H_{ZF}(k)) = x(n) * h(n)$ (linear convolution) over N contiguous samples. The other N sample values of the $2N$-point IDFT are artifacts and are discarded. To do convolution over a KN-sample time series with N-point DFTs, zero padding techniques are used in conjunction with data partitioning methods called overlap and save or overlap and add.

The DFT is used extensively to estimate the transfer function $H(k)$ of a linear system. The transfer function is given by the ratio $H(k) = Y(k)/X(k)$. The shape of the magnitude profile $|H(k)| = \mathrm{SQRT}[H(k)H^*(k)]$ defines the magnitude frequency response of the filter, while $\phi(k) = \arctan\{\mathrm{Im}[H(k)]/\mathrm{Re}[H(k)]\}$ establishes its phase reponse. The derivative of the phase profile, approximated by $\Delta\phi(k) = \phi(k) - \phi(k-1)$, is called the group delay and is often minimized as a design criterion in digital communication/telemetry applications.

The similarity-measuring autocorrelation function, with delay or lag parameter k, can be defined in terms of the DFT as follows:

$$R_x(k) = \frac{1}{N}\sum_{i=0}^{N-1} x(i)x(i+k)$$
$$= \mathrm{IDFT}[X(k)X^*(k)] = \mathrm{IDFT}[|X(k)|^2],$$

provided the zero padding and data partitioning methods required for linear convolution are also employed. The term $X(k)X^*(k) = |X(k)|^2$ is also referred to as the power spectrum of $\{x(n)\}$. The average value, on a per harmonic basis, of a set of power spectra is called a periodogram.

Cross-correlation is similarly defined by

$$R_{xy}(k) = \frac{1}{N}\sum_{i=0}^{N-1} x(i)y(i-k)$$
$$= \mathrm{IDFT}[X(k)Y^*(k)],$$

and the cross-correlation coefficient at delay k, denoted by $\rho_{xy}^{(k)}$, is given by

$$\rho_{xy}^{(k)} = \mathrm{IDFT}[X(k)Y^*(k)]/(|X(0)||Y(0)|)^{1/2}$$

The coherence function is also defined in terms of the DFT and can be used to answer the following questions:

1. Is the system linear?
2. It the initial state of the system zero?
3. Is the system forced by a single input rather than multiple inputs?

In addition, a number of other specialized system-oriented tests have become practical due to the advent of the DFT (i.e., FFT) and DSP procedures. All of the methods have evolved because of the development of efficient and fast DSP computation tools.

VIII. DIGITAL FILTER DESIGN

One of the major branches of study in the digital signal processing field is that of digital filter design and analysis. The design objective is to convert a discrete system, given in terms of its impulse response $[h(n)]$ or transfer function $H(z)$, into hardware or software statement. A filter's hardware design is said to be its architecture. Filter analysis is concerned with issues of speed, precision, and complexity. It was noted earlier that if a digital filter is designed without using signal feedback, it is called a nonrecursive or transversal filter and can be mathematically represented as

$$H(z) = \sum_{n=0}^{N-1} h(n)z^{-n}$$

if the impulse response is given by the N-element time series $\{h(0), h(1), \ldots, h(N-1)\}$. Some special types of

recursive filters also satisfy this condition, but they are rarely seen in practice. Because the filter exhibits a finite impulse response, it is also called an FIR filter. The frequency domain response can be computed by substituting z with $\cdot z = \exp(j\omega)$. In the frequency domain the FIR transform becomes $H(\omega) = |H(\omega)| \exp(\phi(\omega))$, where $|H(\omega)|$ and $\phi(\omega)$ are called the magnitude frequency and phase response, respectively. The principal advantages of FIR filters are twofold: (1) they are simple to analyze since they are defined by a finite linear equation, and (2) they can be designed to have a phase response that satisfies the linear equation $\phi(\omega) = a\omega + b$, where $b = 0$, $\pi/2$, or $-\pi/2$. Linear phase filters are important in applications where nonlinear phase destortion can degrade a database or system performance. Some examples are digital communication line equalization, speech and image processing, and antialiasing filters to precondition the data sent to a DFT. For a FIR filter to achieve linear phase performance, the filter must have a symmetric or antisymmetric finite impulse response. It is also common to find the amplitude profile of a FIR filter overlaid by a symmetric data window such as those introduced in the discussion of DFT. In such cases the original impulse response $\{h(n)\}$ is combined with a window function $\{w(n)\}$ to form a new FIR given by $\{h'(n)\} = \{h(n)w(n)\}$.

An FIR filter can be designed to be frequency selective as well. The four basic frequency selection options are low pass, high pass, bandpass, and bandstop (or bandreject). A popular design procedure approximates the shape of an ideal filter with a Chebyshev polynomial, which is used to define the individual values of $\{h(n)\}$. The ideal filter is assumed to have unit gain in its passband(s) and zero gain in the stopband(s). This design process is now highly automated and is found in many software packages used commercially and in the public domain. Finally, FIR filters also possess a simple but possibly high-order hardware architecture and software code. A direct implementation of the transfer function is suggested in Fig. 8.

If the impulse response of a filter is of infinite duration, the filter is called an infinite impulse response (IIR), and the response is given by

$$H(z) = \sum_{n=0}^{\infty} h(n) z^{-n},$$

which can often be factored into the form

$$H(z) = \left(\sum_{i=0}^{M} b_i z^{-i} \right) \Big/ \left(1 + \sum_{i=1}^{N} a_i z^{-i} \right).$$
$$= N(z)/D(z).$$

The feedforward information paths are characterized by $N(z)$, while the feedback paths are specified by $D(z)$. The solutions to $N(z) = 0$ are called the zeros of filters; the poles are given by $D(z) = 0$. An IIR is said to be bounded if $\sum |h(n)|$, over all n, is finite. If the poles of an IIR are bounded within with the unit circle, then the filter's forced response is also known to remain bounded or stable. Many IIR filters are specified in terms of their optimal analog

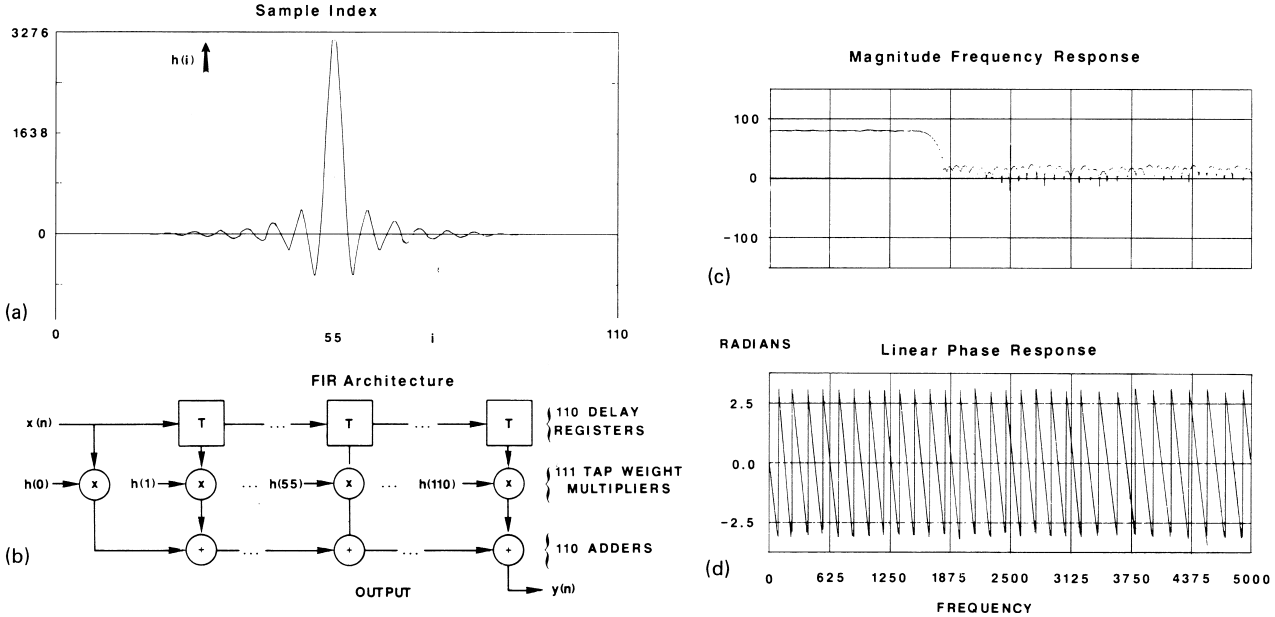


FIGURE 8 Direct implementation of a linear phase FIR lowpass digital filter using shift registers, multipliers, and adders. Phase is plotted between the principal angles $\pm\pi$. (a) Impulse response, (b) architecture, and (c and d) frequency domain performance.

counterparts, which generally take one of the following forms:

Type	Design optimality criteria
Butterworth	Maximally flat
Chebyshev	Equal-ripple passband
Elliptic	Equal-ripple passband and stopband

It is assumed that the ideal filter model has unity gain in the passband and zero elsewhere. The maximally flat criterion refers to the slope (derivative) being zero at a known critical frequency. The equal ripple (or equiripple) is sometimes referred to as a minimax criterion. In a minimax design, the maximum deviation from the ideal model is minimal. The design of a Butterworth, Chebyshev, or elliptic filter $H(s)$ is now a classic study and has been computerized. Examples of classically defined IIRs are presented in Fig. 9.

Infinite impulse response digital filters are designed by reflecting the desired filter parameters [i.e., $H(z)$] into the continuous-frequency domain [i.e., $H(s)$], where the filter equations are well known. This can be accomplished by using one of two procedures, both of which begin with definition of a desired Butterworth, Chebyshev, or elliptic analog prototype filter, denoted $H_{\mathrm{p}}(s)$ in Fig. 10. For prespecified passband and stopband gains A_{p} and A_{a}, the prototype filter satisfies $|H(s)| = A_{\mathrm{p}}$ at $s = j1$ (normalized frequency in radians per second) and has sufficient order to ensure that the gain at the normalized stopband frequency $\Omega'_{\mathrm{a}} = \Omega_{\mathrm{a}}/\Omega_{\mathrm{p}}$ satisfies $|H(\Omega'_{\mathrm{a}})| = A_{\mathrm{a}}$. The frequency terms, denoted Ω in this formula, are in fact the prewarped version of the critical discrete frequencies ω_{p} and ω_{a} as shown in Fig. 3. From that point on, the method proceeds as shown in Table IV (see Fig. 10).

Compared to an FIR, some general conclusions may be drawn about on IIR. First, if magnitude frequency response is to be sharp or abrupt in terms of a transition between passbands and stopbands (called the filter skirt), the IIR is the design of choice. If phase performance is the design objective, an FIR should be chosen. The IIRs are generally of low order ($N \leq 16$), while the FIRs are usually of high order ($N \geq 16$). For the IIR case, the derived transfer function can be factored into a number of recognizable forms. If $H(z) = \Pi\, H_i(z)$, a serial or cascade form is realized. If $H(z) = \Sigma\, H_i(z)$, a parallel design is achieved. The individual subsystems, denoted $H_i(z)$, are normally designed as first- and second-order filters having only real coefficients. These unique structures are called architectures. Other commonly found architectures are the Direct II, canonic, ladder, Gray–Markel, wave, and continued fraction, to name but a few.

Digital filters differ from their discrete counterparts $H(z)$ in that they are designed with digital hardware, arithmetic, and an n-bit data word. If floating-point arithmetic is used, the dynamic range and precision are high but the speed suffers. If speed is an important design issue, fixed-point arithmetic is usually adopted.

A high-speed fixed-point design with a signed $(n+1)$-bit format is subject to quantization errors. Quantization is a process by which a real variable is converted into a finite-precision digital word. Errors introduced by approximating a real variable with its nearest binary-coded decimal (BCD) value manifest themselves as input, coefficient quantization, or arithmetic errors. If a real input is defined over $\pm V$ volts and is presented to an analog-to-digital (ADC) converter for quantization, the quantization step size is given by $Q \leq 1/2^n$ V/bit. The value of the quantization error ε is normally defined as the difference between a real variable X and its nearest binary-valued representation, say X_{B}. That is, $\varepsilon = (X - X_{\mathrm{B}})$ and is usually modeled to be of zero mean for rounding and $\pm Q/2$ for truncation. The error variance in both cases is $Q^2/12 = 2^{-2n}/12$. The error can be reduced by increasing n, but this results in a slower design.

Arithmetic errors are the result of truncation or rounding. For example, the multiplication of two n-bit numbers is a full-precision product of length $2n$ bits. This $2n$-bit product is then rounded or truncated to its nearest n-bit value. The errors associated with this operation can be modeled in the manner used for input or coefficient error analysis. However, in an IIR arithmetic errors can recirculate within the filter through its feedback paths. As a result, once inserted, these errors may never completely leave the filter. The effect of such errors on the frequency response of a Butterworth or Chebyshev low-pass filter, as a function of word length n, is interpreted in Fig. 11. Many of the IIR forms developed over the past several decades represent attempts to reduce the effects of these circulating errors and are called low-sensitivity filters. The error performance of some of the more commonly designed filters is shown in Table V, where complexity can be translated into added cost or slower throughput.

Besides the computational inaccuracies associated with finite-precision effects, other phenomena have been observed. The most important of these is limit cycling, which appears in zero input limit cycling or overflow limit cycling. Zero input limit cycling is a granularity problem in which the natural exponential decay of an impulse response wants to move the output close to zero, but this can never be achieved because of quantizing. For example, if $h(n) = a^n$, then for $a = -1/2$, $h(n) = 1, -1/2, 1/4, -1/8, 1/16, -1/32, \ldots$. However, if a signed 4-bit data word with format $\pm xxx$ is used, the response

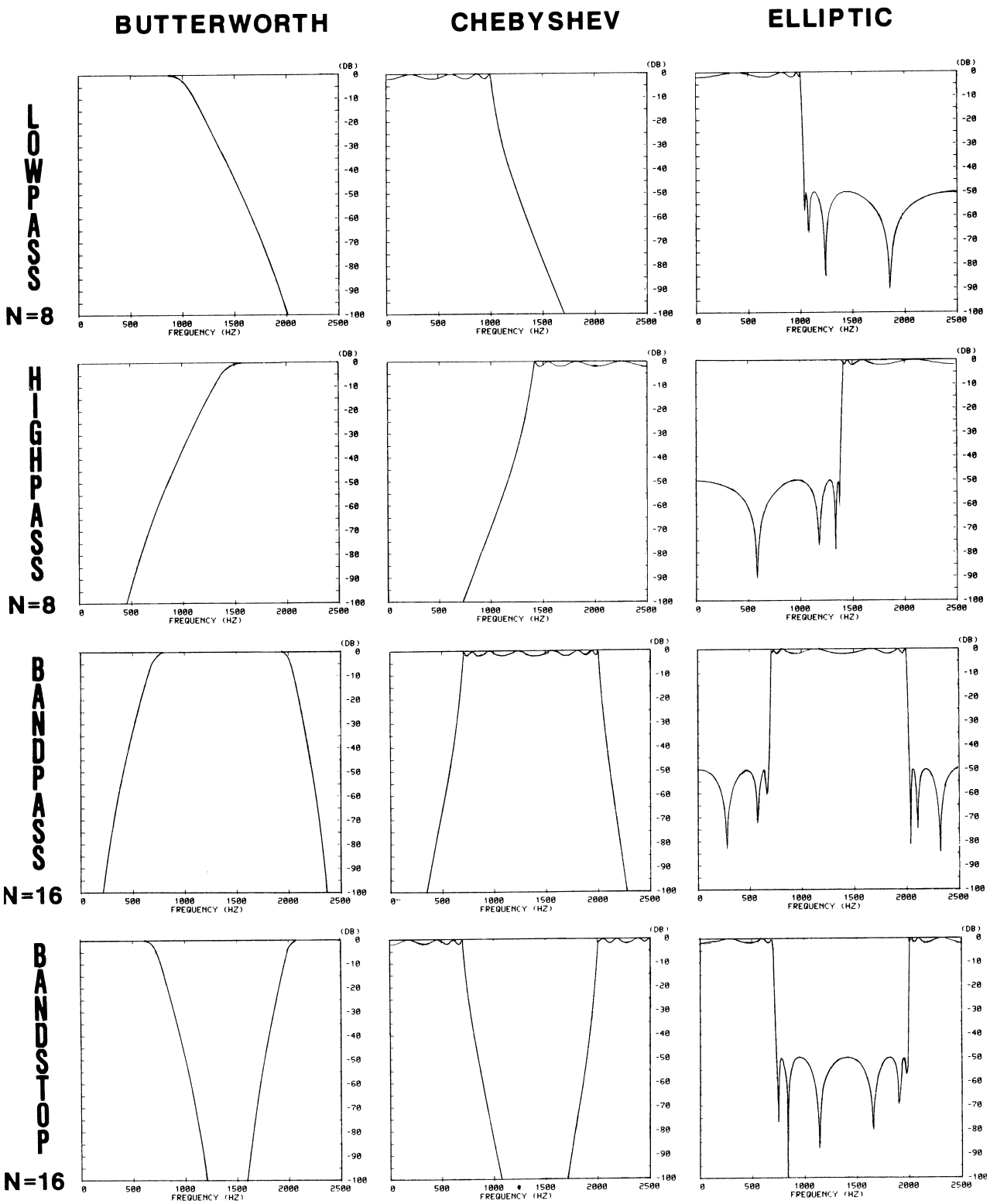
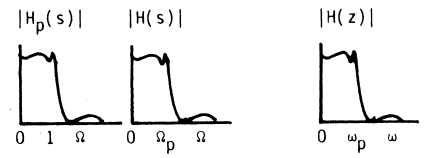
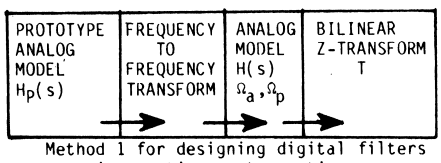
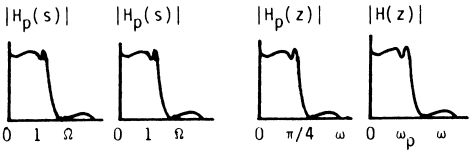


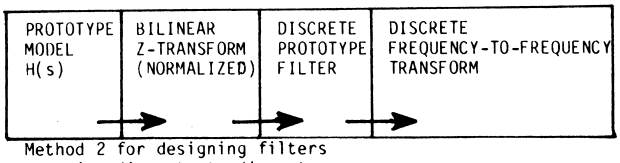
FIGURE 9 Examples of Butterworth, Chebyshev, and elliptic IIR digital filters.



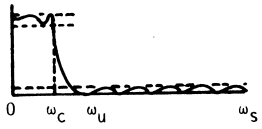
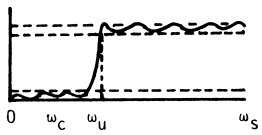
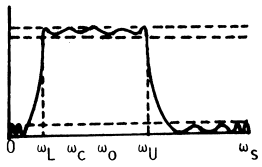
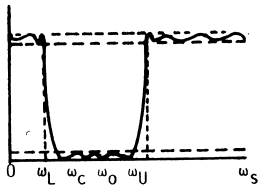


Method 1 for designing digital filters using continuous-to-continuous frequency transforms





Method 2 for designing filters using discrete-to-discrete frequency transforms

Lowpass Model	Continuous-to-Continuous Transform: $H(s) \rightarrow H(\tilde{s})$ #	Discrete-to-Discrete Transform: $H(z) \rightarrow H(\tilde{z})$ %	Discrete frequency response
Lowpass to Lowpass (LP) → (LP)	$\tilde{s} = \frac{s}{\tilde{\Omega}_p}$	$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$ $\alpha = \frac{\sin[(\omega_c - \omega_u)/2]T}{\sin[(\omega_c + \omega_u)/2]T}$	
Lowpass to Highpass (LP) → (HP)	$\tilde{s} = \frac{\tilde{\Omega}_p}{s}$	$\tilde{z}^{-1} = -\frac{z^{-1} + \alpha}{1 + \alpha z^{-1}}$ $z = -\frac{\cos[(\omega_c - \omega_u)/2]T}{\cos[(\omega_c + \omega_u)/2]T}$	
Lowpass to Bandpass (LP) → (BP)	$\tilde{s} = \frac{(s^2 + \tilde{\Omega}_{p_1}\tilde{\Omega}_{p_2})}{s^2(\tilde{\Omega}_{p_1} - \tilde{\Omega}_{p_2})}$	$\tilde{z}^{-1} = -\frac{z^{-2} - \frac{2\alpha k}{k+1} z^{-1} + \frac{k-1}{k+1}}{\frac{k-1}{k+1} z^{-2} - \frac{2\alpha k}{k+1} z^{-1} + 1}$ $\alpha = \cos(\omega_0 T) = \frac{\cos[(\omega_U + \omega_L)/2]T}{\cos[(\omega_U - \omega_L)/2]T}$ $k = \cot\frac{(\omega_U - \omega_L)T}{2} \tan(\frac{\omega_c T}{2})$	
Lowpass to Bandstop (LP) → (BS)	$\tilde{s} = \frac{(s^2(\tilde{\Omega}_{p_2} - \tilde{\Omega}_{p_1}))}{(s^2 + \tilde{\Omega}_{p_1}\tilde{\Omega}_{p_2})}$	$\tilde{z}^{-1} = \frac{z^{-2} - \frac{2\alpha}{k+1} z^{-1} + \frac{k-1}{k+1}}{\frac{k-1}{k+1} z^{-2} - \frac{2\alpha}{k+1} z^{-1} + 1}$ $\alpha = \cos(\omega_0 T) = \frac{\cos[(\omega_U + \omega_L)/2]T}{\cos[(\omega_U - \omega_L)/2]T}$ $k = \tan\frac{(\omega_U - \omega_L)T}{2} \tan(\frac{\omega_c T}{2})$	

$\tilde{\Omega}$ is prewarped frequency

$\tilde{\Omega} = \frac{2}{T} \tan(\frac{\omega T}{2})$

% $\tilde{z}$ is transformed frequency

FIGURE 10 Design procedures for Butterworth, Chebyshev, and elliptic IIR filters.

TABLE IV Design of IIR Digital Filters

Step	Method 1	Step	Method 2
1. Design prototype filter	$H_p(s)$	1. Design prototype filter	$H_p(s)$
2. Convert prototype into a design filter $H(s)$ using a continuous frequency-to-frequency transform	$H(s)$	2. Convert prototype into a bilinear transform prototype	$H_p(z)$
3. Convert $H(s)$ into $H(z)$ using the bilinear z-transform	$H(z)$	3. Convert prototype $H(s)$ into $H(z)$ using discrete frequency-to-frequency transform	$H(z)$

is $y_4(n) = 7/8, -1/2, 1/4, -1/8, 1/8, \ldots, (-1)^n/8, \ldots$. This low-amplitude oscillation of $\pm 1/8$, for $n = 4$, is called zero input limit cycling, and it can result in annoying clicks in speech or telephone applications. Some filter architectures have been developed that trade off additional hardware for suppressed limit cycle behavior. Overflow limit cycling can occur whenever a signal goes into a register overflow condition, which occurs when the signal exceeds the admissible dynamic range of the numbering system. For example, adding 1 to an n-bit maximal binary value $[11 \cdots 1] = 2^n - 1$ (decimal) would send the sum back to zero. Such a rapid and unexpected change in a variable's value may cause bizarre behavior. This effect can be reduced by using saturating arithmetic, which clamps a variable to a maximum (or minimum) admissible value.

Two-dimensional digital FIR and IIR filters can also be designed. If their impulse response $h(n_1, n_2)$ can be factored as $h(n_1, n_2) = h^{[1]}(n_1)h^{[2]}(n_2)$, the filter is said to be separable. Separable filters lend themselves to one-dimensional analysis. For example, if $h(n_1, n_2)$ is separable, the 2D convolution sum can be expressed as

$$\begin{aligned} y(m_1, m_2) &= \sum_{n_1} \sum_{n_2} h^{[1]}(n_1)h^{[2]}(n_2)x(m_1 - n_1, m_2 - n_2) \\ &= \sum_{n_1} h^{[1]}(n_1) \sum_{n_2} h^{[2]}(n_2)x(m_1 - n_1, m_2 - n_2) \\ &= \sum_{n_1} h^{[1]}(n_1)y^{[2]}(m_1 - n_1, m_2). \end{aligned}$$

Finally, the other 1D concepts, such as windows, can be transferred to the 2D case in a straightforward manner.

A filter that has the ability to adjust its coefficients to account for changes in the environment or signal space is called an adaptive filter. Such filters use an algorithm or rule-based decision law to couple measured data back into the filter as a set of superior coefficients. Often this information source is an error signal obtained by actually measuring the difference between a system variable and its adaptively predicted value. By its nature, an adaptive filter must be capable of altering its coefficient set to compensate for change. Therefore, general-purpose multipliers must be used to implement the filter's coefficient. Fixed (nonadaptive) filters, in contrast, often store their coefficients in read-only memory (ROM) and use special wiring techniques to replace expensive general-purpose multipliers with special-purpose ones.

IX. IMPLEMENTATION

Digital processing filters and algorithms are computationally intensive. Arithmetic is performed in a method that is consistent with chosen number systems. In general, three types of numbering systems are used in DSP applications:

1. Weighted (e.g., decimal).
2. Unweighted (e.g., residue).
3. Homomorphic (e.g., logarithmic).

Unsigned weighted numbers take the general form

$$X = \sum_{i=0}^{N-1} a_i r^i, \qquad 0 \le X \le r^N - 1,$$

$$0 \le a_i \le r - 1,$$

where r is the radix and a_i the ith significant digit, with a_0 and a_{N-1} being the least and most significant, respectively. Typical radices found in DSP applications are $r = 2$ (binary), $r = 8$ (octal), $r = 10$ (decimal), and $r = 16$ (hexadecimal). If signed information is required, an additional sign bit is added to the code. In this model, sign-magnitude, 1's complement, or 2's complement fixed-point representations can be defined. The 2's complement fixed-point code is the most popular one because it can be used to add a column of numbers without suffering from otherwise fatal intermittent adder overflow errors, provided the sum has a valid representation.

A fractional number system can be defined by using negative (vs positive) exponents. In either case, additional precision can be gained by adding more digits (i.e., increasing N). However, there is a practical limit to this beyond which the floating-point representation, given by $X = \pm m(x)r^{\pm e(x)}$, where $m(x)$ is a mantissa and $e(x)$ the

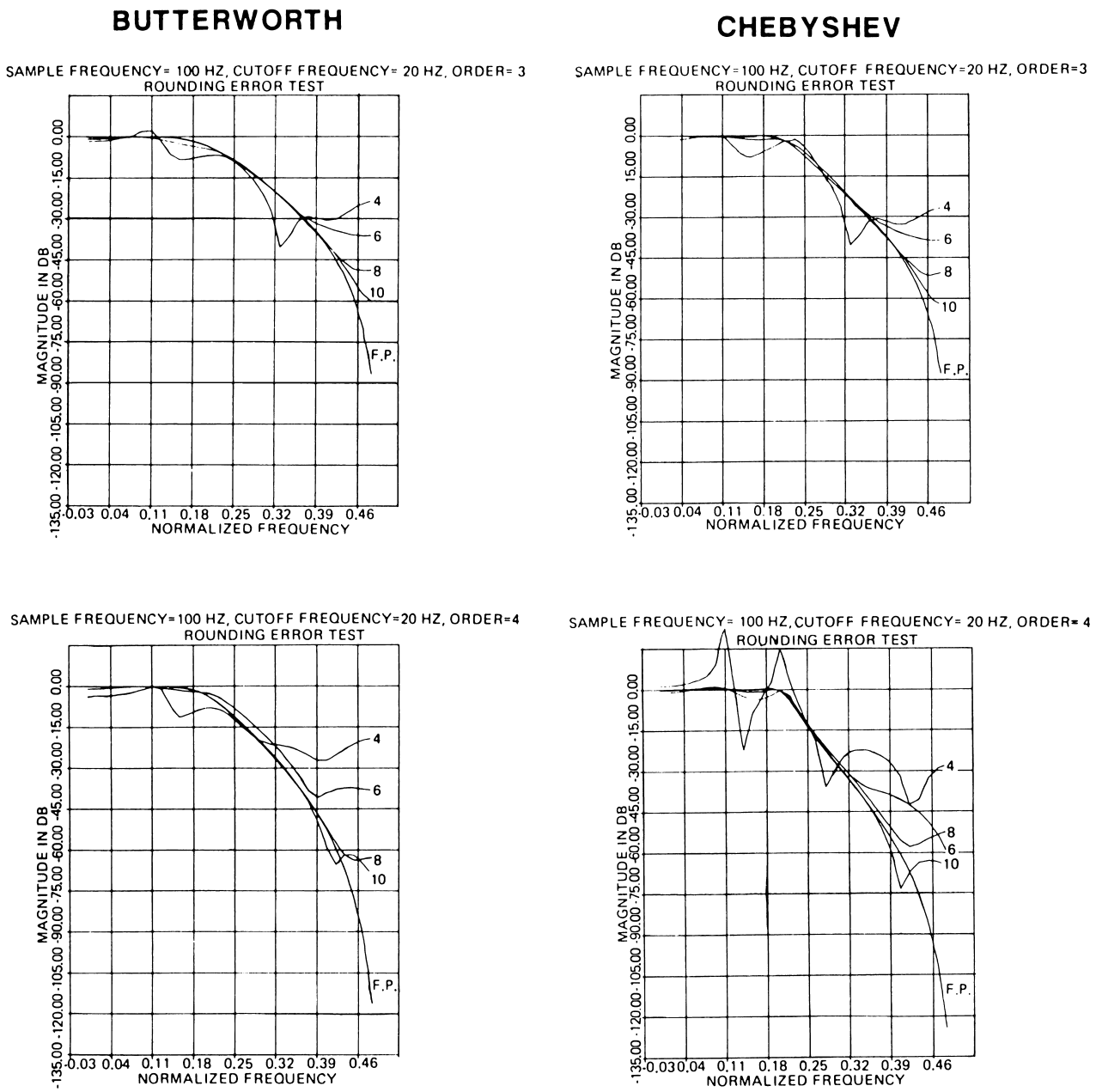


FIGURE 11 Effect of finite arithmetic on the frequency response of digital filters. F.P.: Floating-point response, which is equivalent to an infinite precision filter.

exponent, should be used. Implementation of floating-point arithmetic, though considerably more complex than that of fixed-point arithmetic, is often required in scientific calculations, where large dynamic ranges of data may be encountered. Finally, residue arithmetic (unweighted) is a method of doing carry-free highly parallel arithmetic that dates back to 500 BC. In logarithmic systems (homomorphic) multiplication is performed by adding exponents and is therefore fast. However, addition in this system is slower and more complex than multiplication.

For hardware implementations, the primitive computational unit is the two-input half adder or modulo-2 adder. If this concept is extended to accept a third carry-in input, then a full adder results. A group of full adder cells can be integrated onto a single semiconductor chip to form longer word length adders typically having widths of 4, 8, and 16 bits. If additional hardware is traded off for higher add speeds, then carry lookahead adders or CLAs are used. Otherwise, slower ripple-type adders are chosen because of their simplicity and low cost. Multipliers can

TABLE V Survey of IIR Filter Architectures

Architecture type	Coefficient sensitivity	Round-off hardware error sensitivity	Complexity
Direct II	High	High	Low
Continued fraction	High	High	Low
Cascade	Low	Higher than parallel	Low
Parallel	Low	Lower than parallel	Low
Wave	Low	Good	Significant
Gray–Markel (ladder)	Low	Moderate	Moderate

also be classified in terms of their speed–complexity metrics. The simplest, but slowest, is the shift-add multiplier, which simply emulates the manual multiplication methods learned in elementary school. Fast but more complex units can be designed by using either a multibit technique called Booth's algorithm or a specially modified full adder called a carry save unit. Examples of some commercially available multipliers are summarized in Table VI.

These relatively short word length add and multiply units are conveniently packaged, low in cost, and readily available. Longer word length fixed-point units may be designed by interconnecting the smaller units into arrays. Floating-point units can be similarly designed with separate mantissa and exponent subsystems plus logic to interconnect them.

Dedicated hardware arithmetic units are typically expensive and complex. Slower but more flexible alternatives are micro-, mini-, or maxicomputers. Besides providing arithmetic support, programmable computers are capable of performing operations such as control, storage, and logic. Where speed is a secondary issue, they have become very popular cost-effective DSP design media.

To overcome the latency problem of wide-wordlength fixed-point multipliers, various data compression schemes have been developed. It is assumed that the speed at which the operations of addition, subtraction, multiplication, or division can be performed is inversely related to wordlength. Many of these methods emulate long-standing communication modulation techniques advanced in the 1950s. Most digital filters employ some form of pulse code modulation (PCM), proposed in 1938 by A. H. Reeves. Differential PCM began its commercial life in 1955 with Bell's T1 inverse DPCM operation, denoted $(\text{DPCM})^{-1}$. If the DPCM is modeled as a differential process, reconstruction would require discrete integration of the form $x(n) = y(n) + y(n-1)$. Explicitly, the FIR–DPCM convolution sum can be expressed as

$$[\text{PCM}]\ y(n) = \sum a_i x(n-i):$$

$$[\text{DPCM}]\ y(n) - y(n-1) = \sum a_i [x(n) - x(n-1)],$$

which can be simplified to read $\Delta y(n) = \sum a_i \Delta x(n)$. Under the proper set of conditions, the incremental quantities $\Delta x(j)$ and $\Delta y(j)$ can be amplitude compressed to n_c bits for $n_c \le n$. For example, if $n = 16$ and $n_c = 8$, a filter of conventional architecture would require 16×16 multipliers, compared to 8×8 for the DPCM. However, to ensure that Δx does not exceed its allocated budget of n_c bits, the sampled values of $x(j)$ and $x(j-1)$ must be similar. If they are not, an error-inducing phenomenon called slope overload will occur. This means that the sample rate t_s must often be increased to a value much greater than the original t_s to ensure that a sample $x(j)$ has insufficient time to change to a distinctly different value $x(j+1)$. This is called oversampling. For example, an audio filter having a Nyquist frequency of 25 kHz, with a 16-bit PCM architecture, may prove to be an inferior design based on an 8-bit microprocessor being clocked at 100 kHz.

A very popular form of DPCM is delta modulation (DM), which uses a one-bit $\{+1, -1\}$ DPCM code. Delta modulation can be traced back to a 1946 French patent.

TABLE VI Some Commercially Available Multiplier Units (Circa 1990)

Vendor[a]	Size	Pins	Speed (nsec)	Power (W)	Coding[b]	Algorithm
AMD	2×4	24	25	0.6	2s	Booth
MMI	8×8	40	100	1.0	2s/uns	Modified booth
MMI	16×16	24	800	1.0	2s	Modified booth
TI	4×4	20	50	0.5	uns	Table lookup
TRW	8×8	40	45	1.0	2s	Carry-save
TRW	12×12	64	80	2.0	2s	Carry-save
TRW	16×16	64	100	3.0	2s	Carry-save
TRW	24×24	64	200	4.0	2s	Carry-save

[a] AMD: American Micro Devices; MMI: Monolithic Memories; TI: Texas Instruments.
[b] uns: Unsigned; 2s: 2s complement code.

The DM systems are generally inexpensive but may suffer from severe slope overflow errors. As a result, high oversample rates must be used. These systems also suffer from a phenomenon known as *idle noise*, which is caused by asymmetry in the two current source encoders (i.e., $+1$ and -1) usually found in DM systems and often results in an audible disturbance.

An adaptive DPCM (ADPCM) policy is sometimes used to counter the slope overload problem found in DPCM systems. In an ADPCM system the quantization step size Q is adjusted in real time to reduce the occurrence of slope overload. As a result, the ADPCM system oversample rates are typically lower than those found in DPCM configurations. However, the price paid for those added features is greater system complexity and cost.

Beginning in the mid-1970s, high-speed semiconductor memory-intensive filters became an alternative to the more traditional multiplier-intensive designs. This development was motivated by the fact that fixed-point multiplication can be a rather time-consuming operation compared to other digital computer operations. Further, many digital filters are specified in terms of a set of fixed coefficients. As a result, scaling rather than general multiplication is required. The distinction is that multiplication involves two variables, while scaling involves one. By combining these operations, a DSP unit known as the distributed arithmetic filter was developed. This filter is also referred to by the more confusing title of a bit-slice digital filter. The distributed filter concept is used to realize a fixed-coefficient linear shift-invariant filter of the form

$$y(n) = \sum_{i=0}^{L} a(i)x(n-i) + \sum_{j=1}^{M} b(j)y(n-j),$$

where the $\{a\}$ and $\{b\}$ are known and $L+M+1$ general multiplications are required of traditional designs. If the variables are given an N-bit binary representation $q(t) = \sum_{k=0}^{N-1} q(k,t)2^k$, $k = 0, 1, \ldots, n-1$, where $q(k,t)$ denotes the kth bit of q at time t, then

$$y(n) = \sum_{i=0}^{L}\sum_{k=0}^{N-1} a(i)x(k, n-i)2^k + \sum_{j=1}^{M}\sum_{k=0}^{N-1} b(j)y(k, n-j)2^k.$$

Distributed arithmetic owes its name to the fact that the order of the above addition is *redistributed* as

$$y(n) = \sum_{k=0}^{N-1} 2^k \phi(n;k),$$

$$\phi(n;k) = \left(\sum_{i=0}^{L} a(i)x(k, n-i) + \sum_{j=1}^{M} b(j)y(k, n-j) \right).$$

The function ϕ is defined in terms of the known $\{a\}$ and $\{b\}$ plus the $L+1$ and M binary-valued elements $x(k,s)$ and $y(k,s)$, respectively. These $L+M+1$ binary values can be concatenated into a single $L+M+1$-bit word and used as an address to a ROM that contains the precomputed values of ϕ for all possible 2^{L+M+1} addresses. The sequence of table lookups of ϕ are then simply scaled by 2^k (binary shift) and sent to an adder for accumulation. Since $L+M+1$ multiply and $L+M-1$ add cycles have been replaced by n fast table lookups and $(n-1)$ shift-adds, the distributed filter can often run faster. However, it is nonprogrammable, and to change the filter a new memory table lookup unit (i.e., ϕ) would have to be programmed and installed. For example, if $y(n) = 4x(n) + 3x(n-1) + y(n-1)$, $x(n) = x(n-1) = 1 \rightarrow 0001$ $y(n-1) = 2 \rightarrow 0010$, $n = 3$, then $y(n) = 9$, as shown in the insert at the bottom of this page.

Whether dedicated or programmable arithmetic logic units (ALUs) are used, a hardware DSP system would also consist of memory and peripheral devices. Memory can be broadly classified as read/write random access memory (RAM) or read-only memory (ROM); ROM is often used to store control programs and fixed coefficients, and RAM to store data and variables. Most digital filters also require some sort of peripheral data acquisition support. Analog-to-digital (ADC) and digital-to-analog (DAC) conversion can be quantified in terms of

1. Accuracy: absolute accuracy of conversion relative to a standard voltage or current.
2. Aperture uncertainty: uncertainty in sampling rate.
3. Bandwidth: maximum small-signal 3-dB frequency.
4. Codes: 2's complement, 1's complement, sign-magnitude, etc.
5. Common mode rejection: measure of ability to suppress unwanted noise at input.
6. Conversion time: maximum conversion rate. If a separate sample-and-hold amplifier is used, the conversion rate is the maximum conversion rate of the composite system.
7. Full scale: dynamic range in volts.
8. Glitch: undesirable noise spikes found in A/D output.
9. Linearity: linearity across dynamic range.
10. Offset drift: worst-case variation due to parametric changes.
11. Precision: repeatability of measurement.
12. Quantization error: $\pm(1/2)$ least significant bit (LSB).
13. Resolution: 2^{-n}.

Analog-to-digital and digital-to-analog converters are also classified in terms of word length and speed as summarized in Table VII.

Address $= [x[i,n],$	$x[i,n-1],$	$y[i,n-1]] \rightarrow$	ϕ(Address)	$s \leftarrow s + 2^i\phi$
$i = 0$(LSB) $\rightarrow [1,$	1 ,	$0]$	$(4)1+(3)1+1(0)=7$	$7 \leftarrow 0+2^0(7)$
$i = 1 \rightarrow [0,$	0 ,	$1]$	$(4)0+(3)0+1(1)=1$	$9 \leftarrow 7+2^1(1)$
$i = 2$(MSB) $\rightarrow [0,$	0 ,	$0]$	$(4)0+(3)0+1(0)=0$	$9 \leftarrow 9+2^2(0)$

A useful variation of the standard ADC is the data compression or robust converter. With such a device, data in a large dynamic range can be compressed into a smaller range. A commonly used data compression routine is the mulaw compander, given by

$$y(x) = V \log(1 + \mu x/V)/\log(1 + \mu).$$

Note that for $\mu = 0$ the converter is linear; the value $\mu = 255$ has been found useful in telephone applications.

Digital-to-analog converters (DACs) are of major importance as well. Some DAC units have a three-wire configuration that corrects for a difference between the ground potential of the DAC and the system it is driving. DACs can be extremely useful in supporting test, calibration, and signal synthesis tasks. The monolithic DAC units are basically low in cost and simple to interface. Some standard applications of DACs are in data display, signal generation, and servo-motor control.

Software and computer programs are also important tools in the DSP design process. A number of commercially available DSP system design and analysis packages are marketed to do DFT and digital filter tasks. In developing custom code for a DSP application, software engineering techniques are used that maximize cache memory hits (if cache is present) and reduce the execution latency (delay) of a code.

X. HARDWARE

The technical revolution of the mid 1970s produced the tools needed to affect many high-volume real-time DSP markets. These include medium, large, and very large integrated circuits (MSI, LSI, VLSI) plus DSP metal oxide semi-conductor (MOS) and bipolar devices. The ubiquitous microprocessors, with their increasing capabilities and decreasing costs, can provide trend toward developed single-chip dedicated DSP units. Modern semiconductor technology is being used to implement high-performance real-time dedicated systems for special purposes. Many of these applications are military in nature. Nonmilitary applications, such as computer vision, speech processing, and pattern recognition, have similar performance requirements. Such systems are being developed that make use of pipelined or distributed architectures to increase the speed and capacity of a DSP design. Data flow and systolic structures are also being developed for very high speed application; they typically trade additional complexity and data paths for speed and the simplicity of asynchronous control. For those designing high-volume telecommunications, primitive speech synthesis, and low-order digital filters, a plethora of DSP chips are available. Many of these are hybrid (both analog and digital on a common silicon foundation), charge-coupled device (CCD), or switched capacitance devices. Other semicustom strategies, such as gate arrays, standard cell, or master cell methods, can be used to fill in the gaps between the custom and generic devices.

One of the technology-driven developments in DSP is the DSP chip. Perhaps the most salient characteristic of a DSP chip is its multiplier. Since multipliers normally consume a large amount of chip real estate, their design has been constantly refined and redefined. The early AMI2811 had a slow $12 \times 12 = 16$ multiplier, while a later TMS320 had a $16 \times 16 = 32$ 200-nsec multiplier that occupied about 40% more area. These chips include some amount of on-broad RAM for data storage and ROM for fixed coefficient storage. Since the cost of these chips is very low (tens of dollars), they have opened many new areas for DSP penetration. Many factors, such as speed, cost, performance, software support and programming language, debugging process. Today, a vast array of fixed and floating-point DSP chips are commercially available.

TABLE VII Classification of Analog-to-Digital and Digital-to-Analog Converters

Analog to digital		Digital to analog	
Type	**Conversion rate**	**Type**	**Settling time**
Ultra high speed	>3 MHz	Ultra high speed	<100 nsec
Very high speed	300 kHz ← 3 MHz	Very high speed	100 → 1 μsec
High speed	30 kHz ← 300 kHz	High speed	1 → 10 μsec
Intermediate	3 kHz ← 30 kHz	Intermediate	10 → 100 μsec
Low speed	<3 kHz	Low speed	>100 μsec

Sosme maintain the low-cost uniprocessor architecture of the 1980s, others adopt more modern multiprocessor architecture.

Although the DSP chip concept is exciting, it cannot address all the important DSP problems. The DSP chips are still too slow and imprecise to do some real-time processing. Special processors are often required to do fast "number crunching" DSP operations over a large data base. Sometimes fast arithmetic chips and ASIC cores can be integrated into the design to achieve the needed throughput. In more demanding applications such as speech, geophysics, medicine, and pattern recognition, array processors may be needed. Since most DSP operations are arithmetic intensive, the 10- to 1000-fold increase in speed offered by an array processor can easily justify the substantial cost of these units. Finally, newer generations of DSP hardware designs are concentrating on parallel or pipelined structures. These architectures are tuned to specific applications to achieve maximum throughput.

XI. APPLICATIONS

A. Radar and Sonar Signal Processing

Radar is an important technology that affects all phases of civilian and military aviation as well as space exploration. Whether the radars are pulsed or continuous wave, DSP has become an important design and analysis tool. Because of their speed and reliability, DSP methods have found a welcome home in Doppler and moving target processors. In addition, DSP can serve in target signature analysis applications, where a target's spectral image is analyzed to provide target-type information. The heart of many radar systems is a device called a matched filter. A matched filter, under some stationary signal assumptions, maximizes the output signal-to-noise ratio. From this cleaner database, radar decisions are made. The matched filter convolution operations can be replaced by FFT operations. To achieve very high speeds, pipelined FFTs are often used. These units are generally designed to have a high real-time bandwidth (i.e., transform-per-second rate) and they operate as fast fixed-point or hybrid number system processors rather than slower but more precise floating-point units.

There is some commonality between the principles of sonar and radar. Sonar is based on acoustic waves and appears in both active and passive forms. As with radar, DSP tools can be used to design high-performance correlators and matched filters. Time domain digital filters or FFT methods can be used to implement sonar signal processes. Often the data are collected from an array of sensors, so a plurality of processing units would be integrated into the design.

Since digital shift registers can be used effectively to design delay lines, the physical design of correlators and delay estimators is a manageable task. Because the delayed and manipulated data remain in a digital format, they can be directly passed to a digital processor for high-level analysis. The DSP hardware can also be used to build beam-forming and beam-steering FIR-like filters for sonar and radar applications.

B. Telecommunications and Audio

The telecommunications industry relied almost exclusively on analog devices prior to the 1960s. Systems were designed with amplifiers, capacitors, and inductors. Beginning in the early 1960s, the industry began to convert to a digital format. Digital telecommunication systems have several advantages in terms of speed and bandwidth over their analog ancestors. Early voice grade telecommunication systems used pulse code modulation. In a PCM code each bit represents a fixed amount of signal voltage. If the least significant bit or zeroth bit has a weight of V volts, the nth bit has a weight of 2^nV volts. To achieve recognizable voice quality speech, sampling at rates of 8000 or more samples per second over a 13-bit or greater dynamic range must often be used. To reduce the dynamic range requirement, a logarithmic $\mu - 255$ law data compander can be used to compress speech intelligence into an 8-bit word. Because of the high-volume need for voice-grade telecommunications equipment, the commonly used companding ADCs are often found integrated onto low-cost highly reliable semiconductor chips. Other DSP devices have found their way into the telecommunications industry in the form of signal generators, decoders, and voice synthesizers. By the mid-1970s more than 65 million miles of digital telecommunications circuits were in operation. Because of the high volume of components needed in the telecommunications industry, a number of custom integrated-circuit chips have been developed for use as multiplexers, digital filters, and processors known as echo cancelers. The echo canceler removes the imbalance along a telephone circuit, such as a satellite link, that can cause the phenomenon known as double-talk (an overlapped incoming and outgoing signal or echo). It uses an adjustable or adaptive FIR filter to reconstruct a copy of the unwanted reflected signal, which it subtracts from the outgoing signal before transmission.

Historically, the principal concern of the pre-1970 audio engineer was sound reproduction. This field has now expanded to include digital recording and reproduction, plus digital information storage and transmission. Digital audio is somewhat distinct from digital speech processing in that speech is a telecommunications problem in which the applied criterion of success is intelligibility. In digital audio,

the target is fidelity. In many cases, DSP technology has exceeded the fidelity capability of analog recording and reproduction equipment and is virtually noise-free. Also, digital audio designs have eliminated some of the electromechanical components that deteriorate with age—the phonograph stylus and records, magnetic tape and heads, and so forth. Digital equipment has also been introduced into the recording studio, since it needs much less manual attention and adjustment than analog units and is more versatile.

Audio signal processing begins with an acoustic wave, which is manipulated and possibly archieved. It is returned to an acoustic form for listing. As a result, an ADC and DACs are used at the input–output level. A 16-bit ADC conversion of an analog signal, when played back through a DAC, introduces no detectable noise (distortion) to a human listener. However, the longer the required wordlength, the more expensive the DSP system and storage medium. As a result, data compression techniques are often applied to the digitized database. These include PCM encoding and delta modulation techniques. Between the 1-bit DM protocol and the n-bit PCM is the hybrid differential pulse code modulation, or DPCM. Here the difference between two successive samples is coded as a low-order PCM word and reconstructed by a DM-like incremental addition/subtraction rule. For simple speech-type signals, these methods can achieve a good degree of data compression. For complex music, their use is more difficult.

As analog tapes are played the quality of the recording degrades. Therefore, for studio as well as listener use, storing audio information on digital tape has advantages. Digital storage of audio information may also include some redundant information for use in error correction and detection.

C. Speech Processing

A milestone in speech processing was the development of the vocoder, or voice coder, by Homor Dudley in 1939. His basic speech synthesis model is still in use today. In 1946 the advent of the South spectrograph for short-term speech spectrum display marked the beginning of extensive use of spectral analysis for speech processing. This field has been growing rapidly ever since. Although they are not yet able to perform as well as humans, speech processing algorithms and machines are becoming commercially available.

Human speech is a rather short-term phenomenon that changes dynamically. As a result, its spectral signature is often studied with short-time Fourier transforms. Various windowing and data segmentation methods have been developed to serve this need. Modified transforms such as the Wigner, cepstrum, and smoothed cepstrum can also be used. If $X(k)$ is the DFT of $x(n)$, then the complex cepstrum is given by $x'(n)$, where $x'(n) = \text{IDFT}[X'(k)]$ for $X'(k) = \log X(k) = \log|X(k)| + j \arg X(k)$ $(j = \sqrt{-1})$ and $\arg = \arctan[\text{Im}\, X(k)/\text{Real}\, X(k)]$.

The design of integrated voice/data networks is predicated on minimizing the channel bandwidth (number of bits per second) necessary to transmit voice, while maintaining high speech quality and intelligibility. Typically, the transmission of speech with pulse code modulation requires about 64,000 bits/sec. Specialized adaptive predictive coding (APC) and adaptive transform coding (ATC) techniques can reduce this to about 16,000 bits/sec with no loss in speech quality. Coding techniques below 4000 bits/sec are based on a parametric speech synthesis models such as linear predictive coding (LPC) vocoders and channel vocoders. A further reduction in a bit rate is possible with vector quantization rather than scalar quantization techniques. In the latter, each parameter to be transmitted is quantized independently of all other parameters. In the former, a number of parameters are quantized together as a vector into a multidimensional space. Pattern recognition methods are then used to code and decode the intelligence.

Speech synthesis refers to two distinct types of speech generation by computer and may take the form of stored-parameters algorithms (e.g., the Speak & Spell educational toy using LPC parameters) or wave form coding. The first is text-to-speech synthesis and requires that a computer phonetically "read" a scanned or stored text. The second is speech recognition, which refers to the ability of a machine to recognize or understand human speech. Most commercial and research speech recognition systems today use a pattern-matching approach. In terms of current technology, speaker-dependent continuous speech rates in the high 90th percentile can be achieved only over relatively small vocabularies (tens of words). Large vocabulary, continuous speech, and speaker-independent operation lie somewhere in the future. Finally, speaker recognition systems attempt to recognize people from their voice prints. Present-day techniques can verify a speaker with high accuracy only under controlled conditions.

D. Image Processing

Image processing attempts to extract or modify information found in an image-dependent signal space. The military is a long-time user and promoter of image processing theory, practice, and hardware. In medicine, many diagnostic tools rely heavily on image processors or provide fast analysis of a graphical database. Other image processing applications are in the consumer electronics industry, law enforcement, geophysics, weather prediction,

and the compression, transmission, and reconstruction of television-like signals. In the growing field of robotics, computer vision will play an important role.

Image processing can be partitioned into four subordinate areas known as image restoration, enhancement, coding, and analysis. If one assumes that an image has been degraded due to transmission noise, poor optics, geometric distortion, movement, and so forth, restoration methods are used to repair as much as the damage as possible.

Detectors can convert the optical data found in the image plane to an electronic format, which is then digitized. The point spread function models the optical distortion in the image plane and can be represented as a transfer function. The point spread function shows how much of the information at an image plant "spreads" into neighboring areas in the image plane. Distortion can also be introduced by the recording device, and noise can be added to communication channels.

A basic image restoration task is deblurring. An inverting filter is sought to reduce the effect of noise and produce a facsimile of the source. A difficult nonblurring restoration problem is that of deconvolution, which attempts to recover an image from a set of projections. Often, as in the case of medical X-ray, electromagnetic interference (EMI), and similar scanners, the signal represents the amount of energy detected in a lower-dimensional image plane. Reconstructing the three- or two-dimensional image is analogous to sketching a human face from its shadow. The energy in a two- or one-dimensional plane obtained by illuminating an absorbing three- or two-dimensional object is a projection that can be expressed in terms of its spectral signature. To reconstruct the image with some degree of definition, multiple projections are required. The projection slice theorem states that by collecting a group of one-dimensional projections at various angles a two-dimensional image can be constructed, and from the slices of two-dimensional images, a three-dimensional images can be formed, and so forth. Fourier and other types of fast transforms are generally used to do the data analysis because of the enormous amount of data that must be assimilated to reconstruct an image from its projections.

Enhancement is used to improve the ability to interpret an image database. This is often a subjective measure and relies on human perception. It differs from restoration in that restoration seeks to rebuild the original image from its distorted image, while enhancement seeks to improve on the original. Often, due to transmission bandwidth limitations, an image may lose some of its definition. When assumptions must be made about the shape of an object, enhancement methods return some of the original picture quality.

Image coding is used to compress the representation of an image into as few bits as possible while maintaining a specified degree of image intelligibility. For example, a simple 1024×1024 image, resolved to 16 bits, requires a 2^{24}-bit data field. Such figures can overtax a practical communication channel or cause a memory buffer to overflow when large or multiple images are being processed. Therefore, much attention has been given to image data compression. Fortunately, a typical image contains a large amount of redundant or correlated information. Redundant data can be eliminated by using one of a number of techniques that attempt to make trade-offs between processing speed, memory requirements, and compression capabilities. For example, if most of an image remains unchanged from frame to frame, statistical methods suggest that the data rate can be reduced toward an average of 1 bit/pixel. Furthermore, there is usually a considerable correlation between adjacent pixels (or picture elements), which can also be compressed out of the database. With a DPCM system, a signal is replaced by an estimate of its incremental change. Slowly varying signals are therefore compressed into a small dynamic range. By using this technique, an 8-bit encoded image can often be replaced with a 3-bit DPCM equivalent without a subjective loss of quality. However, if the image consists of sharp high-contrast edges, the DPCM will perform poorly.

More mathematically intense transform-domain image compression methods have been developed that work more directly on the correlation question. However, they can be computationally intense and may degrade picture quality if not correctly applied.

SEE ALSO THE FOLLOWING ARTICLES

COMPUTER ALGORITHMS • DIGITAL FILTERS • DIGITAL SPEECH PROCESSING • IMAGE PROCESSING • NUMERICAL ANALYSIS • RADAR • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, GENERAL • VISION SENSORS FOR ROBOTS • Z-TRANSFORM

BIBLIOGRAPHY

Antonious, A. (1979). "Digital Filters: Analysis and Design," McGraw-Hill, New York.

Blahut, R. (1985). "Fast Algorithms for Digital Signal Processing," Addison-Wesley, Reading, MA.

Oppenheim, A. V., and Schafer, R. (1975). "Digital Signal Processing," Prentice-Hall, Englewood Cliffs, NJ.

Oppenheim, A. V., ed. (1978). "Application of Digital Signal Processing," Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L. R., and Gold, B. (1975). "Theory and Applications of Digital Signal Processing," Prentice-Hall, Englewood, Cliffs, NJ.

Taylor, F. (1983). "Digital Filter Design Handbook," Dekker, New York.



Signal Processing, General

Rao Yarlagadda
Oklahoma State University

John E. Hershey
BDM Corporation

I. Introduction
II. Fundamentals of Signal Processing
III. Technology of signal Processing
IV. Contemporary Signal-Processing Topics
V. Thoughts on the Future

GLOSSARY

A/D converter System that accepts as input an analog signal and an analog reference and provides as output a digital signal.
Convolution Integral (summation) expression for continuous (discrete) signals, generally used to represent an output of a linear system in terms of the input and impulse response of the system.
Deconvolution Procedure that removes the effects of convolution.
Encoder System that takes each quantized sample and represents it by a code word, such as a set of bits.
Quantizer System that produces a distinct set of amplitude levels for a continuous range of amplitude levels of the input signal.

SIGNAL PROCESSING deals with the electrical representation of the processes such as voice and imagery. It comprises those analytical techniques that can be applied to these electrical representations to achieve a variety of objectives. These objectives include noise cleaning for quality enhancement, redundancy removal for compression, and the estimation of basic properties of the signal as a means of analyzing the mechanism producing the signal or the characteristics of the medium or channel through which the signal is conveyed.

I. INTRODUCTION

It is very difficult to define signal processing in a meaningful and succinct way. There is an inbred desire to compare it with statistics and perhaps this is a good place to start. The discipline of signal processing and the science of statistics are very different and yet it is not at all clear what forms a clean dividing line between them. Statisticians typically try to determine which of two or several hypotheses are best supported by the outcome of an experiment. Electrical engineering, within whose bounds signal processing is generally admitted to fall, has greatly benefited from statistical methods. Perhaps the most classic of examples is the decoding of (a priori) equally probable binary signals received through Gaussian noise. The decision region is often represented as in Fig. 1. If we decide that signal S_1 was sent if our detector input, x, is less than the threshold and that signal S_2 was sent if x is equal to or greater than the threshold, then we see that an error is possible as the channel noise can cause x to lie in S_2's region

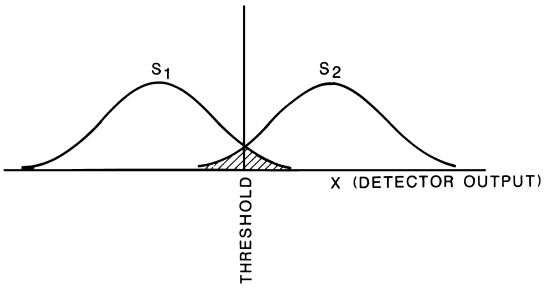


FIGURE 1 The decision regions.

even though S_1 has been sent and vice versa. If the signals are not equally (a priorily) probable, statistical methods that use knowledge of the channel's noise statistics will allow us to pick a threshold that minimizes the probability of incorrect decoding.

Signal processing is a bit more "heuristic" and it is also often a diagnostic process. Instead of trying to decide which of several sources originated a signal, signal analysis will often be used to try to diagnose properties of the processes originating the signals or the characteristics of the medium through which the signals propagate. At the risk of oversimplification, we will say that a statistician will often formulate tests to distinguish between hypotheses *before* examining the data; a signals analyst will often form hypotheses *after* analyzing data.

It has been our experience that it can be extremely helpful to look at a situation both as a statistician and as a signals analyst. It is by a dual viewing that one can perceive the delicate and the aesthetic bridges between various disciplines, the holism that abounds in mathematics.

As a simple example, let us look in two ways at a mistake that has been made over and over throughout the years. The mistake concerns the art of data reduction. In Fig. 2 there are two traces. The top trace is a trace produced by connecting sample value points produced by a white (pseudo)random process. The bottom trace was formed by connecting the points produced by the following "data reduction" routine, applied three times. First, the data were smoothed. Letting x_n be the original data sample values, a new series y_n was formed as

$$y_i = \tfrac{1}{6}(x_i + x_{i-1} + x_{i-2} + x_{i-3} + x_{i-4} + x_{i-5}). \quad (1)$$

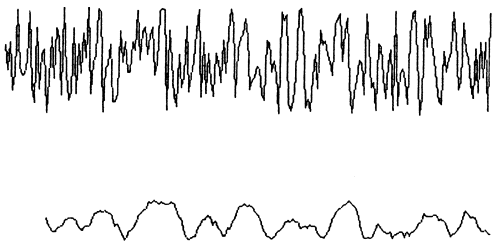


FIGURE 2 Two traces. [Courtesy of R. Kubichek.]

Second, the smoothed data were differenced to form yet another series z_n according to the rule

$$z_i = 2y_i - y_{i-1}. \quad (2)$$

The bottom trace appears to have a periodic component and yet it was derived from a random source. Why is this and why is it so interesting? It is interesting because this is precisely what many scientists have done to some of their data and published a false conclusion based on an apparent periodicity. In advanced statistics, the creation of a false periodicity in this manner is related to the Slutsky–Yule effect that predicts the period of the false oscillation based upon the smoothing and differencing operations. From a signal-processing viewpoint it is immediately clear what has happened. Operation (1) can be viewed as a digital filtering operation that increasingly diminishes the higher frequencies, that is, the higher the frequency the more attenuation. Operation (2), on the other hand, increasingly diminishes the lower frequencies, the lower the frequency the more it is attenuated. Thus the cascaded combination of the two operations results in an overall operation that gives preference to frequencies in a relatively narrow range. Doing the operation three times narrows or "sharpens" this range. Thus when white noise, which contains energy at all frequencies, is applied, the majority of the output signal that will be passed through will consist of energy in this narrow frequency range and the output will resemble an oscillatory signal at the "passband's" center frequency.

II. FUNDAMENTALS OF SIGNAL PROCESSING

A. The Representation of Signals

A signal is a record or history of some physically real value over time or space. A typical signal might be the voltage over time produced by a crystal microphone exposed to the human voice or the two-dimensional array of the values of the point charges produced on a charge-coupled device (CCD) scanning array as it sweeps across an illuminated scene. Signals are represented by a continuous record such as $x(t)$ for our voice voltage or $l(x, y)$ for our luminosity. This is known as the representation of signals in the time or spatial domain. Another, exceptionally important, representation of signals is in the frequency domain. The most common and significant of these representations is that provided by Fourier analysis. Fourier showed that signals could be represented as the weighted sum of sinusoids. For a periodic signal of period T, that is, a signal $x(t)$ for which $x(t) = x(t + T)$ for all t, we can write

$$x(t) = a_0 + 2\sum_{n=1}^{\infty}\left[a_n \cos\left(\frac{2\pi nt}{T}\right) + b_n \sin\left(\frac{2\pi nt}{T}\right)\right] \quad (3)$$

and solve for the coefficients of the sinusoids as follows:

$$a_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\cos\left(\frac{2\pi nt}{T}\right)dt \tag{4a}$$

$$b_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\sin\left(\frac{2\pi nt}{T}\right)dt \tag{4b}$$

Higher frequencies, large values of n in Eq. (4a,b), become increasingly important as the signal becomes closer to a waveform possessing discontinuities. For example, for a periodic triangular waveform that exhibits a discontinuity only in the waveform's derivatives, a_n, b_n decrease quadratically, that is, $a_n, b_n = 0(1/n^2)$; but for a function possessing a discontinuity in the waveform itself, such as a periodic square wave, a_n, b_n decrease only linearly, that is, $a_n, b_n = 0(1/n)$.

If the signal is not periodic (and of finite energy), then the continuous Fourier transform is used:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}\,dt \tag{5a}$$

$$X(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}\,df \tag{5b}$$

Equations (5a,b) are often spoken of as a Fourier transform pair where $X(f)$ is the Fourier transform of $x(t)$ and vice versa.

Just why the Fourier representation is so important is probably due to a confluence of properties. First, the sinusoids are orthogonal, that is, for a periodic signal,

$$\int_{-T/2}^{T/2} \begin{pmatrix}\sin\\ \text{or}\\ \cos\end{pmatrix}\left(\frac{2\pi mt}{T}\right)\begin{pmatrix}\sin\\ \text{or}\\ \cos\end{pmatrix}\left(\frac{2\pi nt}{T}\right)dt$$

$$= \begin{cases} T/2, & m = n \\ 0, & m \neq n. \end{cases} \tag{6}$$

Second, the Fourier transform is a linear operation, that is, the Fourier transform of $c_1x_1(t) + c_2x_2(t)$ is $c_1X_1(f) + c_2X_2(f)$, where c_1 and c_2 are arbitrary constants. Third, the Fourier transform can be easily scaled in time or frequency. The Fourier transform of $x(ct)$ is $(1/|c|)/X(f/c)$, where c is an arbitrary scale factor. Fourth, the magnitude of the Fourier transform is invariant under time displacement, it undergoes only a phase shift, that is, the Fourier transform of $x(t - t_0)$ is $X(f)\exp(-j2\pi ft_0)$. Conversely, a frequency shift is such that the Fourier transform of $X(f - f_0)$ is $x(t)\exp(j2\pi f_0t)$. The time (or spatial) shift property is important in pattern recognition where it is known that a particular waveform is present somewhere in time (or space). The converse frequency shift property is at the heart of many communications processes such as heterodyning. Fifth, sinusoids "occur" naturally as the modes of familiar vibrations and other natural physical processes.

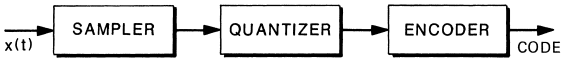


FIGURE 3 The conversion of a continuous signal to a series of discrete samples.

Signals are often represented by a series of discrete samples $\{x(n)\}$ for a one-dimensional signal such as our voice waveform or $\{l(i, j)\}$ for an image. Making the conversion from $x(t) \rightarrow \{x(n)\}$ involves the three steps depicted in Fig. 3. The SAMPLER measures the exact value of $x(t)$ at, for our discussion, periodic intervals, $t = kT$, where T is the sampling interval and k is an integer. Note that the sampling frequency is $f_s = 1/T$. The QUANTIZER maps the exact sampled value of $x(t)$ into one of L quantization levels. The ENCODER assigns a unique code word to each of the L quantization levels and accordingly to each of the sampled values of $x(t)$.

As an example of this process of converting a continuous signal to a series of discrete samples, consider that we have the time trace shown in Fig. 4. Let us now sample $x(t)$ at the 12 points indicated by heavy dots on the time axis as shown in Fig. 5. Let the sampled values be

$$\{2.6, 2.4, 1.6, -1.9, -2.4, -2.6, -2.6, -1.8, -0.4, 1.4, 2.4, 2.4\}. \tag{7}$$

Now let us quantize the values by using the quantizer depicted in Fig. 6. After quantization, the set of values in Eq. (7) have been quantized to the set

$$\{3, 2, 2, -2, -2, -3, -3, -2, 0, 1, 2, 2\}. \tag{8}$$

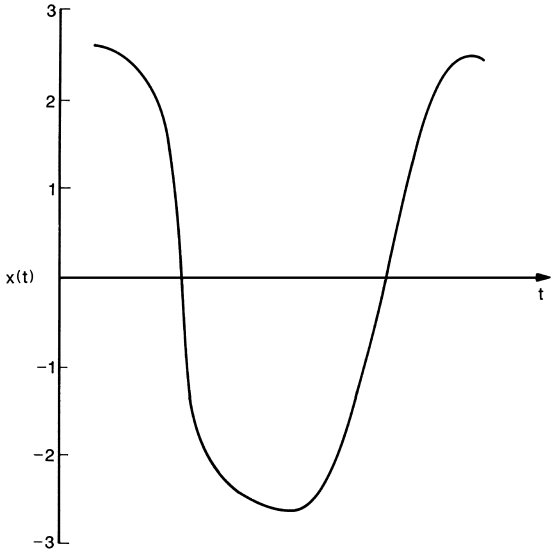


FIGURE 4 An example of $x(t)$.

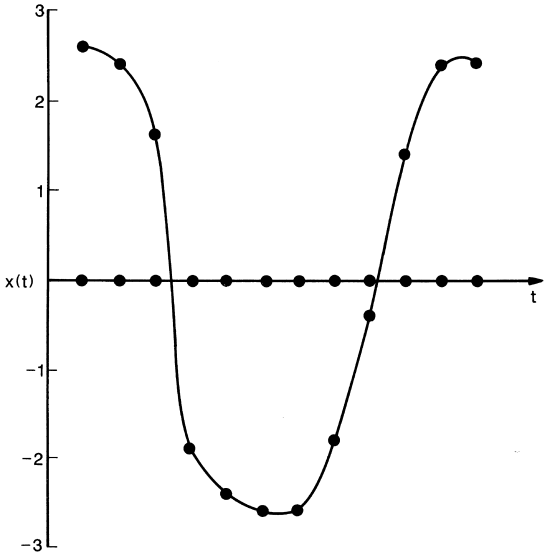


FIGURE 5 Sampling the continuous signal.

The ENCODER for our example is described by the operation specified in Table I. The twelve quantized values in Eq. (8) thus become the series of code words in Eq. (9):

$$\{110, 101, 101, 001, 001, 000, 000, 001, 011, 100, 101, 101\} \quad (9)$$

or the "continuous" bit stream 110101101001001000000001011100101101.

The process we have just gone through is known in general as analog-to-digital (A/D) conversion. There are a number of important considerations in A/D conversion. Starting with the SAMPLING process we encounter the first of these and that is the question "How fast must I sample—what is the minimum acceptable value of f_s?" The answer to this basic question incorporates the name of the man who answered it: Nyquist's rule. Nyquist's rule specifies that sampling must occur at a frequency greater than twice the highest significant frequency component of the waveform to be sampled. Failure to comport

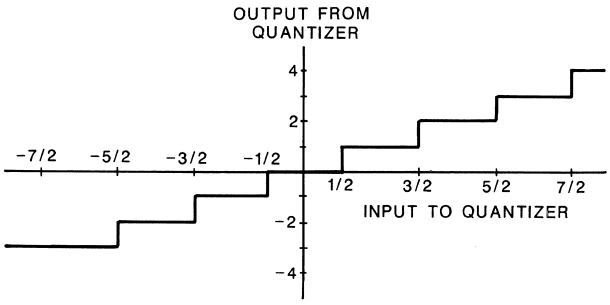


FIGURE 6 Quantizer.

TABLE I ENCODER's Assignment of Code Words to QUANTIZER Outputs

Quantizer output	Code word
−3	000
−2	001
−1	010
0	011
1	100
2	101
3	110
4	111

to Nyquist's rule can lead to "aliasing." This undesirable phenomenon "folds" high-frequency components into the lower frequencies. To see this, consider Fig. 7 where we depict a sinusoid of frequency 1 Hz sampled at 1.33 Hz, a frequency less than the required Nyquist sampling frequency of greater than 2 Hz. In Fig. 7 we have sketched a smooth waveform through the sample points. The reconstructed waveform, that is, the waveform that appears to be present from the "undersampling" is a sinusoid of period 3 s or frequency 0.333 Hz. Thus, a relatively high-frequency component of 1 Hz has been aliased to a frequency component of one-third its frequency. For this reason, one usually prefaces the input of the SAMPLER with a low-pass filter (LPF) that removes any frequencies above that allowed by Nyquist's rule.

The second item we must consider is the effect of the QUANTIZER. Clearly the QUANTIZER will be important if we wish to be able to preserve the fidelity of the sample signal. First, the QUANTIZER's dynamic range must accommodate the dynamic range of the sampled signal most, if not all, of the time. Second, the QUANTIZER must have a sufficient number of quantizing levels to capture the requisite nuances of the signals. The QUANTIZER of Fig. 6 of our example is an example of a multilevel midtread uniform quantizer. It is uniform because the spacing of the quantization levels is a constant. It is called midtread (as opposed to midriser) because zero and the region about zero is handled as a distinct level. Many quantization schemes have been developed and employed over the years. In digital telephony, nonuniform quantizers called companders are extremely important. For digital signal processing, however, the uniform quantizer is preeminent. A very useful "rule of thumb" is that the

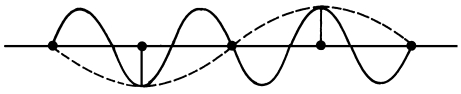


FIGURE 7 Aliasing error.

signal-to-quantizing noise ratio for a uniform quantizer increases about 6 dB for each doubling of the number of quantization levels.

The ENCODER is also important. The code word table should be matched to the signal processing data handling architecture. Note that the digital stream that results from sampling an analog signal at a rate of f_s samples per second and quantizing to L different levels and encoding those levels with a table such as Table I will be $f_s \cdot \log_2 L$ bits per second.

B. Windowing

Because measurements of a particular signal or image are limited to samples of finite scope in time or space, it is necessary to consider the effects of this constraint. The study of these effects falls under the general discipline of "windowing." The term is apt as what we are allowed is a "window" into our process instead of a complete time or spatial record.

The simplest window is the rectangular, sometimes called "boxcar," window. This window is the familiar rect function:

$$\text{rect}(x) = \begin{cases} 1, & |x| \le T/2 \\ 0, & |x| > T/2. \end{cases} \tag{10}$$

What is the effect of such a function? Let us suppose that the process we wish to examine is a pure sinusoid of frequency f_0 that has existed since $t = -\infty$ and will exist to $t = +\infty$. All we see of the sinusoid is a sampling from $t = -T/2$ to $t = +T/2$.

We can calculate the magnitude of the Fourier transform of the short segment of a sinusoid that results from overlaying the rect window on a sinusoid. All we need do is take the Fourier transform of the rect function and then use the frequency shifting property of the Fourier transform. The result is a Fourier transform whose magnitude is as depicted in Fig. 8. From Fig. 8 we see that our windowing has given rise to a rather broadband signal. This may be quite undesirable in many situations and has led many researchers to look for more appropriate window functions for various problems. A triangular window, for example, deemphasizes the beginning and ending values of the sampled signal and, for our sinusoid, leads to a Fourier transform that is much "tighter" about f_0 than is the case for the rect window.

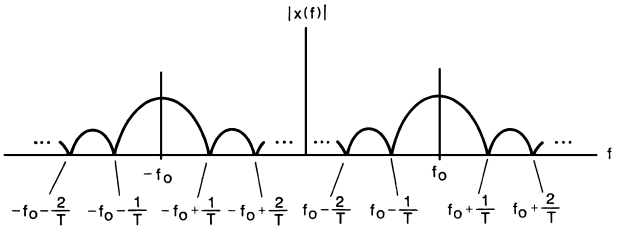


FIGURE 8 The magnitude of the Fourier transform of a short segment of a sinusoid.

C. The Discrete Fourier Transform

In Eq. (5), we have introduced the continuous Fourier transform and its inverse transform. Analytical computations of Eq. (5) are possible for only a very few signals. In most situations $x(t)$ cannot be expressed in an analytical form and it may even be represented by a curve or perhaps its values are known at equal intervals. These reasons point to the direction of the discrete representation of Eq. (5) and its evaluation. Before we can write the discrete representation, we should note a few aspects. First, the signal $x(t)$ needs to be represented at discrete values, say, at $t = nT$, where T is selected according to Nyquist's rule. Second, the integral of $x(t)$ over the small interval $nT \le t < (n+1)T$ can be approximated by

$$\int_{nT}^{(n+1)T} x(t)e^{-j2\pi ft}\,dt \simeq Tx(nT)e^{-j2\pi fnT}, \tag{11}$$

which is usually referred to as rectangular integration where we have assumed that the integrand does not change much in the interval. Finally, we sum over all the intervals. Now,

$$X(f) = T\sum_{n=-\infty}^{\infty} x(nT)e^{-j2\pi fnT}. \tag{12}$$

Most signals are assumed to be nonzero only for a finite interval. Assuming $x(t) = 0$ for $t < 0$ and $t \ge NT$, where N is some integer, we can write Eq. (12) in the form

$$X(f) = T\sum_{n=0}^{N-1} x(nT)e^{-j2\pi fnT} \tag{13}$$

which, incidentally, is periodic with period $1/T$. Sampling $X(f)$ at intervals of $1/NT$, we can write

$$X\left(k\frac{1}{NT}\right) = T\sum_{n=0}^{N-1} x(nT)e^{-j(2\pi/N)kn} \qquad k = 0, 1, \ldots, N-1. \tag{14}$$

Since T and $1/NT$ are constants, they are usually omitted and the discrete Fourier transform (DFT) is expressed by

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j(2\pi/N)kn} \qquad k = 0, 1, \ldots, N-1. \tag{15}$$

Using the analysis similar to the above, we can approximate Eq. (5b) by

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k)e^{j(2\pi/N)kn} \qquad n = 0, 1, \ldots, N-1, \tag{16}$$

which is the inverse discrete Fourier transform (IDFT).

From Eqs. (15) and (16), it is clear that they can be interpreted as linear transformations and they convert one set of data from one domain to another (time to frequency and vice versa). The computation of $X(k)$ in (15) takes N^2 complex multiplications and additions, which gets large when N is large. Fast algorithms have been developed. An algorithm, that is considered to be classic now is given by Cooley and Tukey and is applicable when N is a power of 2. This algorithm and its variations are sometimes referred to as fast Fourier transforms (or FFTs). An outline of this algorithm is presented next.

The idea in this algorithm is to express an N-point DFT by two $N/2$-point DFTs. That is, if

$$A(k) = \sum_{i=0}^{(N/2)-1} x(2i)e^{[-j2\pi/(N/2)]ik}$$
$$k = 0, 1, 2, \ldots, (N/2)-1 \tag{17}$$
$$B(k) = \sum_{i=0}^{(N/2)-1} x(2i+1)e^{[-j2\pi/(N/2)]ik}$$
$$k = 0, 1, 2, \ldots, (N/2)-1, \tag{18}$$

then

$$X(k) = A(k) + B(k)e^{-j(2\pi/N)k} \qquad k = 0, 1, \ldots, N-1. \tag{19}$$

From Eqs. (17) and (18), we see that $A(k)$ and $B(k)$ are DFTs of even and odd samples, respectively. Note also that $A(k)$ and $B(k)$ are periodic with period $(N/2)$. Since an N-point DFT is reduced to two $N/2$-point DFTs, it can be seen that this will reduce the number of operations approximately by half. Successive application of this idea will reduce the number of operations to $N \log_2 N$ as compared to N^2 by direct computation. It is clear that the DFT and the IDFT have the same general forms and, therefore, the same algorithm can be used for both cases with slight modification.

Before we leave this topic, we want to point out certain facets of the FFT. First, the basic FFT algorithm is computed in $(L+1)$ stages corresponding to 2^L data points. Let us denote the sequence of complex numbers resulting from the mth stage of computation as $X_m(i)$, $m = 0, 1, \ldots, L$. Further, let us define the input sequence in the bit-reversed order and define

$$X_0(i) = x(\hat{\imath}),$$

where i is obtained from $\hat{\imath}$ by bit reversion. For example, if $N = 8$, $\hat{\imath} = 3$, then

$$(\hat{\imath})_2 = 011,$$

and reversing the bits, we have

$$(i)_2 = 110 = (6)_{10}.$$

The heart of the FFT computation is in the so-called "butterfly operation," which simply corresponds to computing the output array for the $(m+1)$st stage via

$$X_{m+1}(p) = X_m(p) + e^{-j(2\pi/N)r}X_m(q)$$
$$X_{m+1}(q) = X_m(p) - e^{-j(2\pi/N)r}X_m(q),$$

where r depends on the stage and indices of the variables p and q. Various other fast algorithms are available for computing the DFT when N is a power of two, and for arbitrary N.

The discrete Fourier transform requires multiplications that are computationally expensive compared to additions. A transform that uses only additions and subtractions is the Walsh–Hadamard transform, which is discussed next for N a power of two.

D. Walsh–Hadamard Transform

The standard class of Walsh–Hadamard matrices of order $N = 2^L$ can be generated from the recursive relation

$$H_k = \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}, \qquad k = 1, 2, 3, \ldots, \tag{20}$$

with $H_0 = 1$. For example, when $N = 4$, we have

$$H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \tag{21}$$

The Hadamard transform can be defined by

$$\mathbf{y} = H_L\mathbf{x}, \tag{22}$$

where $\mathbf{x}$ is an N-dimensional data vector and $\mathbf{y}$ is an N-dimensional transform vector. The inverse transform is given by

$$\mathbf{x} = (1/N)H_L\mathbf{y}. \tag{23}$$

Various fast algorithms are available to implement Eqs. (22) or (23). The algorithm that is intuitive uses the basic recursive structure

$$H_k = \begin{bmatrix} I & I \\ I & -I \end{bmatrix}\begin{bmatrix} H_{k-1} & 0 \\ 0 & H_{k-1} \end{bmatrix}, \tag{24}$$

where I is an identity matrix of order 2^{k-1} and the second algorithm uses the relation

$$H_L = A^L, \tag{25a}$$

where

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & & 1 & -1 \end{bmatrix} \tag{25b}$$

Using either one of the structures in Eq. (24) or Eq. (25) in Eq. (22) allows for the computation of the Walsh–Hadamard transform (WHT) in $N \log_2 N$ additions and/or subtractions.

The uses of DFT and WHT are too numerous to mention. First and foremost, the DFT coefficients give the normalized spectral values of the signal. Other uses include applications in data compression, convolution, deconvolution, etc. In the data compression application, for example, we could transmit transformed data rather than raw data, as many of the DFT coefficients (or WHT coefficients) are close to zero and therefore only a portion of the coefficients need to be sent. Next let us consider convolution and correlation.

E. Convolution and Correlation

Before we consider these, there are some concepts that are of interest in system theory. In our study we assume that a system is defined as a unique transformation that maps an input signal, $x(t)$, into an output signal, $y(t)$. Let us denote this by

$$y(t) = T[x(t)]. \tag{26}$$

There are three classes of systems that are of interest to us. These can be defined by putting constraints on the transformation $T[\cdot]$.

1. Linearity

Given the responses $y_1(t)$ and $y_2(t)$ corresponding to the respective inputs $x_1(t)$ and $x_2(t)$, then a system is linear if and only if

$$\begin{aligned} T[a_1x_1(t) + a_2x_2(t)] &= a_1T[x_1(t)] + a_2T[x_2(t)] \\ &= a_1y_1(t) + a_2y_2(t), \end{aligned} \tag{27}$$

where a_1 and a_2 are arbitrary constants. An interesting example of linearity is the delay operation

$$r(t) = T[x(t)] = x(t - t_0), \tag{28}$$

where t_0 is some positive constant. This is a linear operation since

$$T[a_1x_1(t) + a_2x_2(t)] = a_1x_1(t - t_0) + a_2x_2(t - t_0). \tag{29}$$

Others include those of the derivative, integration, etc. An example of a nonlinear operation is

$$r(t) = T[x(t)] = x^2(t). \tag{30}$$

2. Time Invariance

Let $y(t)$ be the response to the input $x(t)$. Then $y(t - t_0)$ is the response to the input $x(t - t_0)$ if the system is time invariant.

Note that this requires that the response be the same anytime you apply the input except that the response is shifted according to the time of the input.

3. Causality

A causal system is a system for which the output for any time t_0 depends on the inputs for $t \le t_0$ only. That is, the response does not depend on the future inputs and it relies only on the past and present inputs.

The above concepts allow for the characterization of a linear system. One function that is important to us is the impulse function, $\delta(t)$, which is defined in terms of the process

$$\int_{-\infty}^{\infty} x(t)\delta(t)\,dt = x(0), \tag{31}$$

where $x(t)$ is any test function that is continuous at $t = 0$. Equation (31) is a special case of the so-called sifting property

$$\int_{-\infty}^{\infty} x(t)\delta(t - t_0)\,dt = x(t_0), \tag{32}$$

where again $x(t)$ is assumed to be continuous at $t = t_0$. An example of an impulse function is shown in Fig. 9, where we have

$$\delta(t) = \lim_{\varepsilon \to 0} \delta_\varepsilon(t). \tag{33}$$

Intuitively we see that any function having unit area and zero width in the limit as some parameter approaches zero is a suitable representation of $\delta(t)$. A dynamite explosion is a good approximation, for example, for an impulse input, which is used in seismic exploration. In the following we will consider exclusively linear time invariant

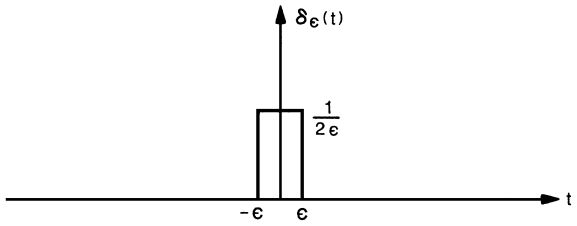


FIGURE 9 An example of an impulse function, as $\varepsilon \to 0$.

systems. The impulse response, $h(t)$, of a linear time invariant system is defined to be the response of the system to an impulse at $t = 0$. By the linear time invariant properties, we can see that the response of a linear system to the input

$$x(t) = \sum_{n=1}^{N} a_n \delta(t - t_n) \tag{34}$$

is

$$y(t) = \sum_{n=1}^{N} a_n h(t - t_n). \tag{35}$$

Using the sifting property in Eq. (32), we can write

$$x(t) = \int_{-\infty}^{\infty} x(t')\delta(t - t')\, dt'. \tag{36}$$

Using the rectangular integration, we can approximate Eq. (36) by

$$x(t) = \sum_{n=-N_1}^{N_2} x(n\Delta T)\delta(t - n\Delta T)\Delta T, \tag{37}$$

where ΔT is some small increment time. The response $y(t)$ in Eq. (35) is given by

$$y(t) = \sum_{n=-N_1}^{N_2} x(n\Delta T)h(t - n\Delta T)\Delta T. \tag{38}$$

As $\Delta T \to 0$, $n\Delta T$ approaches the continuous variable t', the sum in Eq. (38) becomes an integral, and we have

$$y(t) = \int_{-\infty}^{\infty} x(t')h(t - t')\, dt' = \int_{-\infty}^{\infty} h(t')x(t - t')\, dt', \tag{39}$$

which is usually referred to as a convolution integral. The second integral in Eq. (39) is obtained from the first by redefining the variable. The convolution is an important relation, which, in words, says that the response of an arbitrary input is related to the impulse response via Eq. (39). Equation (39) is symbolically written in the form

$$y(t) = x(t) * h(t) = h(t) * x(t), \tag{40}$$

where $(*)$ represents convolution.

The computation of the convolution in Eq. (39) can be computed analytically only if $x(t)$ and $h(t)$ are known analytically and can be integrated. Another approach is to use transform theory. That is, if the Fourier transforms of $x(t)$ and $h(t)$ are expressed as

$$F[x(t)] = X(f) \tag{41}$$

and

$$F[h(t)] = H(f), \tag{42}$$

then we can show that

$$Y(f) = F[y(t)] = H(f)X(f) \tag{43}$$

and

$$Y(t) = \int_{-\infty}^{\infty} H(f)X(f)e^{j2\pi ft}\, df. \tag{44}$$

The transform of $h(t)$, $H(f)$, is usually referred to as a transfer function relating the input and the output. Transfer functions play a major role in communication theory, control systems, and signal processing. We will consider the discrete convolution later. Next, let us consider the concepts of correlation for the aperiodic case.

The process of correlation is useful for comparing two signals and it measures the similarity between one signal and a time-delayed version of a second signal. The cross-correlation of $x(t)$ and $g(t)$ is defined by

$$R_{gx}(\tau) = \int_{-\infty}^{\infty} g(t)x(t + \tau)\, dt, \tag{45}$$

which is a function of τ, the amount of time displacement. When $g(t) = x(t)$, $R_{gg}(\tau)$ is referred to as autocorrelation. Correlation functions are useful in many areas. For example, if $x(t) = g(t - t_0)$, then $R_{gx}(t)$ will be a maximum when $\tau = t_0$ indicating a method for measuring the delay. It is of interest that the cross-correlation in Eq. (45) can be expressed in terms of convolution in Eq. (40). It can be shown that

$$R_{\bar{g}x}(\tau) = g(\tau) * x(\tau), \tag{46}$$

where $\bar{g}(t) = g(-t)$. That is, the cross-correlation of $g(-t)$ with $x(t)$ will result in the convolution of $g(t)$ with $x(t)$.

Transforms can be used to compute correlations also. For example, if $F[x(t)] = X(f)$, $F[g(t)] = G(f)$, and $F[R_{gx}(\tau)] = S_{gx}(f)$, then

$$S_{gx} = G^*(f)X(f), \tag{47}$$

where $G^*(f)$ is the complex conjugate of $G(f)$. Computing the inverse transfer of Eq. (47) will give the cross-correlation function. Next, let us consider the discrete computation of convolutions and correlations.

4. Computation by Discrete Methods

From Eq. (38), we can write

$$y(k\Delta T) = \sum_{n=-\infty}^{\infty} x(n\Delta T)h[(k-n)\Delta T]\Delta T.$$

Suppressing ΔT for simplicity, that is, $\Delta T = 1$, we have

$$y(k) = \sum_{n=-\infty}^{\infty} x(n)h(k-n) = \sum_{n=-\infty}^{\infty} h(n)x(k-n), \quad (48)$$

which is usually referred to as discrete convolution and $h(n)$, the unit sample response. In a similar manner, we can write the correlation in Eq. (45) in discrete form

$$R_{gx}(k) = \sum_{n=-\infty}^{\infty} g(n)x(n+k). \quad (49)$$

The computational procedures are very similar for both convolution and correlation and, therefore, we will concentrate on convolutions. It should be pointed out that Eqs. (48) and (49) could have been derived for either a continuous or a discrete system.

For physical systems $h(n) = 0, n < 0$, that is, they are causal, and furthermore we will assume for practical reasons that $h(n)$ and $x(n)$ are time limited. Let

$$x(n) = 0, \qquad \text{for } n < 0 \text{ and for } n \geq L, \quad (50a)$$

$$h(n) = 0, \qquad \text{for } n < 0 \text{ and for } n \geq N. \quad (50b)$$

It is clear that the lower limit on n for $x(n)$ could be different and the analysis is very similar. With Eq. (50) in Eq. (48), we can rewrite

$$y(k) = \sum_{n=0}^{L-1} x(n)h(k-n) = \sum_{n=0}^{N-1} h(n)x(k-n). \quad (51)$$

Noting the lengths of $x(n)$ and $h(n)$ are L and N, respectively, we see that the output $y(k)$ is time limited and is nonzero for $0 \leq k \leq N + L - 2$. We stress that the convolution of two sequences will result in a sequence longer than either one. Let us illustrate these ideas by a simple example.

Example

Let

$$x(0) = 1, x(1) = 1 \quad \text{and} \quad x(n) = 0, \quad \text{for} \quad n < 0 \quad \text{and} \quad n > 1,$$

and

$$h(0) = 1, h(1) = -1 \quad \text{and} \quad h(n) = 0, \quad \text{for} \quad n < 0 \quad \text{and} \quad n > 1.$$

From Eq. (49), we can write

$$\begin{aligned} y(n) &= 0, \quad n < 0 \\ y(0) &= x(0)\,h(0) + x(1)h(-1) = x(0)h(0) \\ y(1) &= x(0)\,h(1) + x(1)h(0) \\ y(2) &= x(0)\,h(2) \\ y(n) &= 0, \quad n > (N + L - 2) = 2. \end{aligned} \quad (52)$$

The equations in (52) can be written in a matrix form for $0 \leq n \leq 2$

$$\begin{bmatrix} y(0) \\ y(1) \\ y(2) \end{bmatrix} = \begin{bmatrix} h(0) & 0 \\ h(1) & h(0) \\ 0 & h(1) \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \end{bmatrix}. \quad (53a)$$

In symbolic form, Eq. (53a) can be written as

$$\mathbf{y} = H\mathbf{x}. \quad (53b)$$

Note the nice structure of the coefficient matrix in Eq. (53a). This structure can be generalized. Incidentally, the correlation equations have the same general structure as in Eq. (53). Direct computation of Eq. (51) is computationally intensive, especially when the number of output data, $N + L - 1$, is large. For this case, the FFT is generally used to compute convolutions and correlations. This is discussed next. For an effective use of the FFT, first, let M be a power of two and

$$M \geq N + L - 1. \quad (54)$$

Second, pad $x(n)$ and $h(n)$ with 0's. That is, the new $x(n)$, $h(n)$, and $y(n)$ are

$$\begin{aligned} x'(n) &= \begin{cases} x(n), & 0 \leq n \leq L-1 \\ 0, & L \leq n < M \end{cases} \\ h'(n) &= \begin{cases} h(n), & 0 \leq n \leq N-1 \\ 0, & N \leq n < M \end{cases} \\ y'(n) &= \begin{cases} y(n), & 0 \leq n \leq N+L-2 \\ 0, & N+L-1 \leq n < M. \end{cases} \end{aligned} \quad (55)$$

Third, compute the DFTs of $x'(n)$ and $h'(n)$ using the FFT. That is,

$$\begin{aligned} X'(k) &= \text{DFT}[x'(n)] \\ H'(k) &= \text{DFT}[h'(n)]. \end{aligned} \quad (56)$$

Fourth, compute the products, that is,

$$Y'(k) = H'(k)X'(k). \quad (57)$$

Fifth, compute the IDFT of $Y'(k)$, that is,

$$y'(n) = \text{IDFT}[Y'(k)] \quad (58)$$

and delete the last $M-(N+L-1)$ points from $y'(n)$ to obtain $y(n)$. Note the similarity between the discrete computation of $y(n)$ via the DFT and the computation of $y(t)$ via Fourier transforms. We should note the padding in Eq. (55) and the constraint in Eq. (54). Similar approaches can be given for computing the correlations.

III. TECHNOLOGY OF SIGNAL PROCESSING

The technology of signal processing, like so many other technologies, got a primal boost during World War II. It was indeed quite remarkable what the scientists and engineers accomplished with what today seems bulky and crude. Figure 10 is a picture of one of the most important instruments of those days, the sound spectrograph. The sound spectrograph depicted presented a hard copy, visual display of the energy present in different frequency ranges over a time record of modest length. The vertical axis was devoted to frequency which typically spanned the range from 0 to 3+ kHz. The horizontal axis was used for time; a typical trace was of about 2 s in duration. The sound spectrograph of the early 1940s technology thus exhibited an analysis capability whose time-bandwidth product was of the order of 10^4.

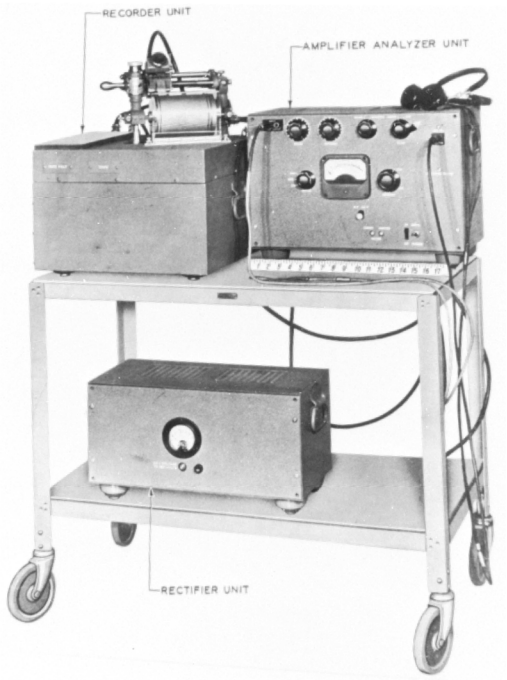


FIGURE 10 Sound spectrograph. [From speech and facsimile scrambling and decoding, summary Technical Report of Division 13, NDRC Vol. 3, Washington, D.C., 1946.]

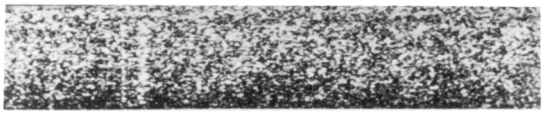


FIGURE 11 Spectrograph—thermal noise. [From speech and facsimile scrambling and decoding, summary Technical Report of Division 13, NDRC Vol. 3, Washington, D.C., 1946.]

Figures 11–15 are typical of some research sound spectrographs produced in the early 1940s. Figure 11 is the sound spectrograph associated with thermal noise. Note the random distribution of energy over frequency and time. Figure 12 is the sound spectrograph associated with a telephone bell. Note the significant low-frequency component and the alternating pattern in the two upper frequency components probably brought about by the clapper alternating strokes between two slightly different bells. Figure 13 is the sound spectrograph produced by crumpling paper. Notice that the crumpling action results in broadband, noise-like impulses. Figure 14 is a sound spectrograph of piano music. Note the multitone components. Figure 15 is a sound spectrograph of normal human speech of the words "one, two, three, four, five, six." Note the heavy deposits of energy in the time varying "bands." These bands are termed the formants of speech. They are the resonances excited in the oral and nasal cavities of the vocal tract by the periodic pulse train generated at the glottis for voiced sounds such as the vowels.

Today, of course, most signal processing is done by high-speed electronic digital computational devices. Because so many of the signal-processing tasks share computational "modules," such as matrix-processing tasks, a class of computational engines has gained prominance in contemporary signal processing. These machines are the array processors (not "arrays" of processors that will be discussed later) that are serial, "pipelined," computational devices that are extremely efficient when dealing with certain repetitive, deterministic (nondecision, nonbranching) computational code such as is written for manipulation of matrices or numerical arrays. One such device is pictured in Fig. 16. The device is capable of performing 100 million floating point computations per second when it is able to run at maximum efficiency. These array processors or, as they are sometimes called, numerical coprocessors, are very popular and widespread and also somewhat difficult to compare without actual "hands-on," problem-specific experience.

Comparisons between any competing entities are extremely difficult to make the more complex the entities are. To attack the problem, the computational community has informally and collectively devised a potpourri of comparison guidelines usually termed "benchmarks." The motivation for developing benchmarks is innate to our

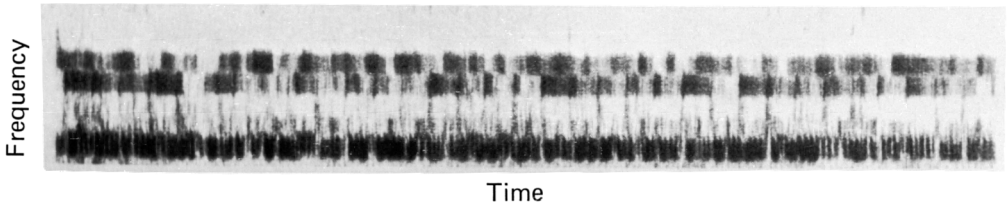


FIGURE 12 Spectrograph—telephone bell. [From speech and facsimile scrambling and decoding, summary Technical Report of Division 13, NDRC Vol. 3, Washington, D.C., 1946.]

culture and also it is a general desire of management personnel to "dedimensionalize" complex problems to reduce vast and complex issues to simple figures-of-merit.

Benchmarking has been done on (1) hardware, such as fast coprocessors, (2) software, such as operating systems, and (3) entire computational systems. The desire to concoct the perfect and universal razor is an attractive intellectual pursuit. It may, however, be fruitless as the true "bottom line" is the efficacy of a piece of hardware, software, or system in solving a particular—hypothetical—problem.

One can be fully secure in the suspicion that a vendor will disseminate the performance of those benchmarks that reveal the product in the most favorable light.

The following terms are widely used when comparing processors. They have some meaning for serial machines but become inadequate, and even misleading, for highly parallel structures.

1. MIPS: millions of instructions per second; generally the reciprocal of the cycle time.
2. LOPS: logical operations per second; as many machines are capable of performing register-to-register logic instructions, LOPS can be thought of as the product of MIPS times the machine's word length.
3. (M)FLOPS: (mega, or millions of) floating point operations per second; generally estimated by the inverse of the time (in microseconds) required to execute a floating point multiplication.

Looking a bit more closely at serial machines, it is often the case that a problem will be sized in time as the ratio of millions of floating point operations required to solve the problem divided by the MFLOPS capability of the machine. That the time required is often more than three times longer than predicted is surprising and disheartening but eminently reasonable and can be easily seen from the following analysis from Parkinson and Liddell.

Let t_c be the cycle, the time required to fetch an operand (from memory or write a result to memory), t_i be the time required to perform integer arithmetic, t_L be the time required to perform a logical test, and t_{FM} be the time required to perform a floating point multiply.

Let us examine what is involved in the following program:

$$\begin{aligned} &\text{DO} \quad 1\ I = 1, N \\ 1 \quad &Z(I) = X(I)^{*}Y(I) \end{aligned}$$

Each pass through the loop requires that we

1. fetch two operands from memory: $X(I)$ and $Y(I)$,
2. floating point multiply the operands,
3. store the result, $Z(I)$, in memory,
4. increment an integer counter $I \leftarrow I + 1$,
5. logically test the counter to see if it exceeds N.

The time, T, required to perform (1)–(5) is

$$T = 3t_c + t_i + t_L + t_{FM}. \tag{59}$$

If the computation of T is limited by the floating point computation, then we might opt to invest in a high-speed floating point coprocessor, that is, we would seemingly be attracted to a device that had a high $1/t_{FM}$, a large number of megaflops. But what really happens? Let us assume

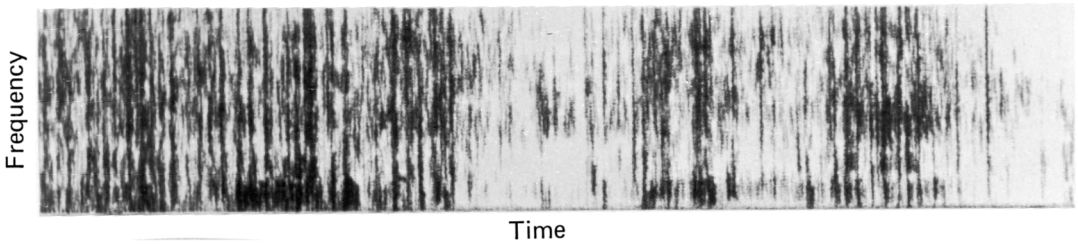


FIGURE 13 Spectrograph—crumpling paper. [From speech and facsimile scrambling and decoding, summary Technical Report of Division 13, NDRC Vol. 3, Washington, D.C., 1946.]

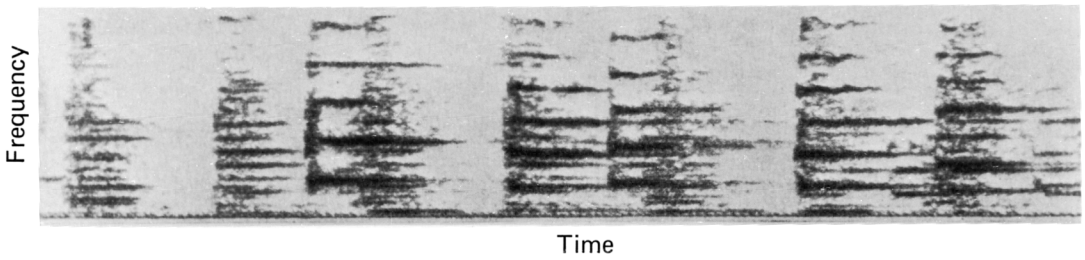


FIGURE 14 Spectrograph—piano music. [From speech and facsimile scrambling and decoding, summary Technical Report of Division 13, NDRC Vol. 3, Washington, D.C., 1946.]

that the t_{FM} before speedup is t_{FM1} and t_{FM2} after speedup. We would be tempted to believe that we had reduced our program's running time by

$$t_{FM2}/t_{FM1}. \tag{60}$$

Unfortunately this is not so. Let us designate T_1 as the value of T in Eq. (59) before floating point speedup and T_2 as the value of T after floating point speedup. The true ratio of times is then

$$\frac{T_2}{T_1} = \frac{3t_c + t_i + t_L + t_{FM2}}{3t_c + t_i + t_L + t_{FM1}}. \tag{61}$$

If, for example $t_c = 200$ ns, $t_i = 100$ ns, $t_L = 100$ ns, and $t_{FM1} = 1200$ ns, $t_{FM2} = 200$ ns, the time is reduced by only 50%, not by one-sixth which is the ratio between t_{FM2} and t_{FM1}. It is the "overhead" that accompanies floating point operations that diminishes the ratio. This concept of overhead is crucial and will be seen again when we consider parallel architectures.

Note that even if t_{FM2} were to vanish (become zero) in Eq. (61), the results would be scarcely more dramatic. This analysis led one group of researchers to opine that "MFLOPS seems to be more popularly used as a measure for systems whose floating point hardware is so powerful that the run time is dominated by all the other terms!"

But computers are not the only way to process signals. A discovery by Lord Rayleigh, published in 1885, has great significance to contemporary signal processing—especially that processing that must be done in "real time." Rayleigh's work concerned the existence of a mode of elastic wave propagation, at most a few wavelengths deep, along the surface of a solid. These Rayleigh, or surface acoustic, waves (SAW) can be used to process analog signals over impressive bandwidths and dynamic ranges.

The usefulness of applying SAW techniques stems from the simple observation that acoustic waves travel much more slowly than electromagnetic waves in free space; typically on the order of one one-hundred thousandth the velocity. Thus, an entire pulse train, which, if sent as an electromagnetic wave through space, might stretch for a great distance, if converted by a transducer to a surface acoustic wave might be entirely captured on a small SAW device effectively serving as a delay line.

By inserting transducers along the SAW medium, it is possible to perform all sorts of signal-processing functions and this is where SAW devices have become so important. The transducer elements can be laid along the SAW to provide many useful impulse responses. Thus the SAW can perform convolution, or cross-correlation, and matched filters, both fixed and programmable, are one of the most important applications of SAW technology that is to be found fulfilling many signal-processing tasks especially in communications and radar processing.

Another separate technology that is gaining in importance for signal processing is optical processing. Optical processing will become increasingly important for those special cases that require enormous time-bandwidth products and real-time computational throughputs of relatively limited precision. Optical processing is so attractive because (1) the computations are done at the speed of light and (2) in parallel.

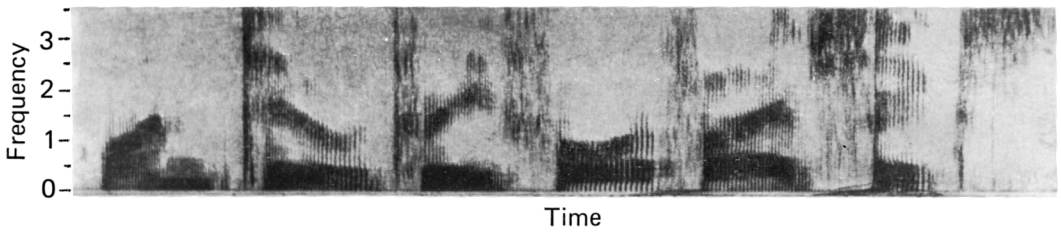


FIGURE 15 Spectrograph—normal human speech: one, two, three, four, five, six. [From speech and facsimile scrambling and decoding, summary Technical Report of Division 13, NDRC Vol. 3, Washington, D.C., 1946.]

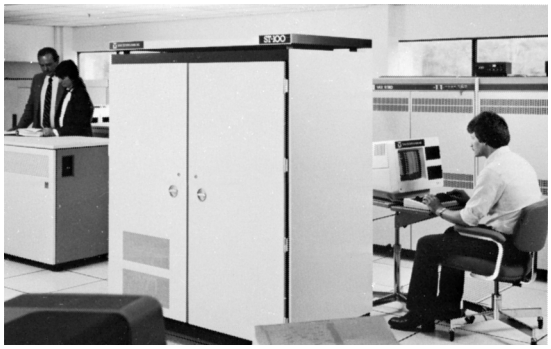


FIGURE 16 Array processor. [Courtesy of Star Technologies, Inc.]

The cornerstone of optical processing is due to the remarkable property that a lens can aid in performing a Fourier transform. In a very cogent and straightforward manner, Stark has shown the relevant mathematics that we summarize. Assume that a monochromatic plane wave (such as produced by a laser) is incident upon an object laying in the wave's plane. The diagram in Fig. 17 depicts the object, the lens, and the position of the lens' back focal plane (f spatial units to the right of the lens' center). We need consider a one-dimensional object only. (The extension to two dimensions is straightforward.) The x-coordinate is assigned to measure spatial distance in the object's dimension and the u-coordinate for spatial frequency in the back focal plane. Assuming that the object's transmittance function is $t(x)$, the free space wavelength of the plane wave is λ, and that the lens aperture has a pupil function $P()$, Stark shows that the complex amplitude of the optical radiation at the back focal plane is proportional to

$$\frac{1}{\sqrt{j\lambda f}} \exp\left[j\pi \left(1 - \frac{d}{f}\right) u^2 \lambda f \right] \cdot \int_{-\infty}^{\infty} t(x) P(x + \lambda u d) e^{-j2\pi u x}\, dx. \qquad (62)$$

Stark points out the following:

1. If the object is set at the lens' front focal plane, that is, at $d = f$, then Eq. (62) becomes the traditional Fourier transform integral. Note that the pupil function serves as a spatial "window" in the same manner as the time windows we looked at earlier.

2. If the object is set against the lens, that is, $d = 0$, then the pupil will have no effect if the object's dimensions do not exceed the lens' aperture. The phase factor (preceding the integral) will not vanish as it did for the case in which $d = f$, however, for this case the pupil will not affect the transform by attenuating the higher frequencies, a phenomenon known as "vignetting." The phase factor is of no consequence, of course, if one wishes to measure only the power spectrum, the square of the magnitude of Eq. (62).

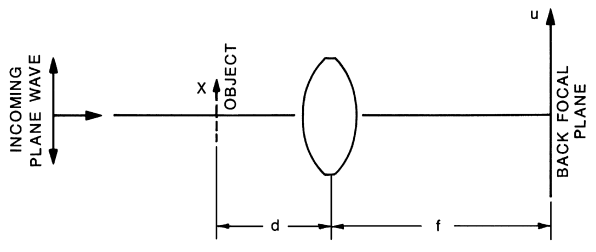


FIGURE 17 A lens performing a Fourier transform.

The possibilities for signal processing implied by Eq. (62) are enormous. A simple and rather ingenious application was a device, constructed by a company for sale, that examined needles to determine if they were sufficiently sharp. As the degree of sharpness was reflected in the needle's surface shape, it was reasonable to expect that the Fourier transform of the shape could serve as a discriminant. This is exactly what the company's device did. The needles were passed by a lens that aided in taking the Fourier transform. Needles that were too dull exhibited a Fourier spectrum that had larger low spatial frequency components than needles that were sufficiently sharp.

More sophisticated and powerful concepts are also possible with relatively simple optics. For example, by placing a "Fourier transform mask"—an object whose transmittance function is itself the Fourier transform—of a scene at the back focal plane of Fig. 17 and then passing the transmitted light through another lens, it is possible to perform the convolution, or, equivalently, the cross-correlation of two scenes. An immediate application of this would be to pattern recognition in such uses as scene template matching.

We should remember that it is often through the synergy that is obtained on melding two distinct and separate technologies that we realize the true quantum leaps in any discipline. The future seems especially promising for the interplay of acoustics and optics, and much progress has already been made, but the state of the art is such that electronic computers will most likely dominate as the tool of choice for signal processing for at least another decade.

IV. CONTEMPORARY SIGNAL-PROCESSING TOPICS

A. Deconvolution

In an earlier section we discussed the convolution of two signals, $x(n)$, the input signal, and $h(n)$, the unit sample response. In most measuring systems, the data are recovered

through a measuring device. That means what we have is not the actual data, but the data convolved with the unit sample response of the measuring device. Obviously, to get the true data we need to get rid of the measuring device's unit sample response. This process is usually referred to as deconvolution. In some applications, for example, speech processing, $x(n)$ is assumed to be known and the vocal tract transfer function needs to be determined. In either case, we need to deconvolve the received data to find $x(n)$ or $h(n)$. This is discussed next for the case of solving for $x(n)$. The other case can be dealt with similarly.

The deconvolution problem can best be stated in terms of matrices. Given the following system of equations [see Eq. (51)]

$$\mathbf{y} = H\mathbf{x}, \tag{63}$$

where $\mathbf{y}$ is an $M = (N + L - 1)$-dimensional vector, H a rectangular matrix of dimension $(N + L - 1) \times L$ and of rank L, and $\mathbf{x}$ an L-dimensional vector, find $\mathbf{x}$ with the assumptions that H and $\mathbf{y}$ are known. If there is a unique solution, then it is very easy to solve, as we need to solve a set of L independent equations. Unfortunately, the problem is not that simple in practice, since the received vector is not the same as $\mathbf{y}$ as it is corrupted by noise. That is, the received vector, say $\mathbf{d}$, is

$$\mathbf{d} = \mathbf{y} + \mathbf{n}, \tag{64}$$

where $\mathbf{n}$ is some noise vector. In most applications the noise is assumed to be additive. Now the problem is to minimize the difference between the desired vector $\mathbf{y}$ and the received vector $\mathbf{d}$ in some sense. The most common measure is an L_p measure, which is discussed next.

Let y_i and d_i be the ith respective components of the M-dimensional vectors $\mathbf{y}$ and $\mathbf{d}$. Furthermore, let

$$E = \sum_{i=0}^{M-1} |d_i - y_i|^p , \tag{65}$$

where $1 \le p \le 2$ is of most interest, and the error E is referred to as L_p based. In vector notation, we have

$$E = \|\mathbf{d} - \mathbf{y}\|_p = \|\mathbf{d} - H\mathbf{x}\|_p. \tag{66}$$

When $p = 2$, the solution is referred to as the least squares or L_2 solution, and for other p's, we simply identify them as L_p solutions. The least-squares solution or the L_2 solution is the simplest to compute and can be determined by setting

$$dE/dx_i = 0, \qquad i = 0, 1, 2, \ldots, L - 1 \tag{67}$$

and solving for the unknowns. The solution can be expressed in a nice compact form, and is

$$\mathbf{x} = (H^{\mathrm{T}}H)^{-1}H^{\mathrm{T}}\mathbf{d}. \tag{68}$$

The matrix $(H^T H)^{-1}H^T$ is sometimes referred to as the pseudo-inverse of the rectangular matrix H. Next let us consider other L_p solutions.

Unfortunately there is no simple solution as in Eq. (68) for other values of p. There are however, iterative techniques for obtaining L_p solutions. The convergence of these iterative algorithms is assured for $1 \le |p| \le 2$ and we assume that in the following.

1. Iteratively Reweighted Least-Squares (IRLS) Algorithm

The IRLS algorithm solves a set of overdetermined system of equations

$$H\mathbf{x} = \mathbf{d} \tag{69}$$

by iteratively computing

$$\mathbf{x}(k + 1) = [H^{\mathrm{T}}W(k)H]^{-1}H^{\mathrm{T}}\mathrm{W(k)}\mathbf{d}, \tag{70}$$

where $W(k)$ is a diagonal matrix with its diagonal entries, $W_{ii}(k)$, given by

$$W_{ii}(k) = \begin{cases} |r_i(k)|^{p-2}, & |r_i(k)| > \varepsilon \\ \varepsilon^{p-2}, & |r_i(k)| \le \varepsilon, \end{cases} \tag{71}$$

where $r_i(k)$ corresponds to the ith entry in the residual vector

$$\mathbf{r}(k) = \mathbf{d} - H\mathbf{x}(k) \tag{72}$$

and ε is some small positive number. The IRLS algorithm is computationally intensive, as we need to find the inverse of the matrix $[H^{\mathrm{T}}W(k)H]$ of order L at each stage. A simpler algorithm is the residual steepest descent algorithm. This is discussed next.

2. Residual Steepest Descent (RSD) Algorithm

The RSD algorithm solves Eq. (69) by iteratively computing

$$\mathbf{x}(k + 1) = \mathbf{x}(k) - \Delta_k(H^{\mathrm{T}}H)^{-1}H^{\mathrm{T}}\gamma(k), \tag{73}$$

where

$$\gamma(k) = \mathrm{col}[\gamma_1(k)\gamma_2(k)\cdots\gamma_M(k)] \tag{74}$$

with

$$\gamma_i(k) = |(Hx(k) - \mathbf{d})_i|^{p-1}\mathrm{sgn}(H\mathbf{x}(k) - \mathbf{d})_i, \tag{75}$$

where $\mathrm{sgn}(t) = +1(-1)$ if $t > 0$ $(t < 0)$. When $t = 0$, one can arbitrarily choose $\mathrm{sgn}(t)$ to be either $+1$ or -1. The scale factor Δ_k in Eq. (73) is determined by minimizing

$$E(k + 1) = \left\| -\mathbf{d} + H\mathbf{x}(k) - \Delta_k H(H^{\mathrm{T}}H)^{-1}H^{\mathrm{T}}\gamma(k) \right\|_p \tag{76}$$

with respect to Δ_k in the L_p sense. Note that Δ_k is a scalar and fast convergence can be obtained by using the IRLS algorithm to compute Δ_k.

In the case of the L_p-based deconvolution, $1 \le p \le 2$, values of p close to 1 produce deconvolution filters with less sensitivity to aberrant noise than when p is close to 2. L_1 deconvolution is generally considered more robust than L_2 deconvolution.

B. Image Transforms and Their Use in Image Coding

Let P be a square matrix of dimension $2^n \times 2^n$. We will define the transform of P as

$$B = A^{\mathrm{T}} P A, \tag{77}$$

where A is any nonsingular matrix. Normally the matrix describing an image, the image matrix, consists of integers with the integers identifying the image brightness of the picture elements or pixels (small regions of the image). The integers range from 0 to $2^L - 1$, where L depends on the system. For our analysis, we will assume that the entries of P, p_{ij}, are from the real field, even though the entries in P are integers. Some examples of A in Eq. (77) are the discrete Fourier transform matrix, Hadamard matrix, etc. In the following, we will use an interesting transform operator that will be useful for image coding. Consider

$$A = \alpha I + (1 - \alpha) H_n, \tag{78}$$

where H_n is the Walsh–Hadamard matrix defined earlier, $0 \le \alpha \le 1$ and $\alpha \ne 1/2$. In a way the transform in Eq. (77) with A given in Eq. (78) is a hybrid transform between the image P and the transformed image. Substituting Eq. (78) in Eq. (77) and expanding, we have

$$B = \alpha^2 P + (1 - \alpha)^2 H_n P H_n + \alpha(1 - \alpha)(P H_n + H_n P). \tag{79}$$

There are three parts in this expression. The first term is the weighted original image. The second term is the weighted Hadamard transform. The third term is a cross-product term. Interestingly, the inverse of A in Eq. (78) is ($\alpha \ne 1/2$)

$$A^{-1} = \frac{1}{(2\alpha - 1)}[\alpha I - (1 - \alpha) H_n], \tag{80}$$

indicating that the inverse transform can be obtained with almost the same structure as the forward transform.

For a significant range of α below unity, the sense of the scene is interpretable, the features are recognizable, but there is a large Hadamard component. These two properties allow for a unique approach to error correction and/or image compression. Since the scene is interpretable for a range of α, we refer to this transform as the naturalness preserving transform (NPT). Assume that a section of B in Eq. (79) is lost or discarded. We can recover P by minimizing the texture noise. Texture is hard to define. It is the feature of an image or subimage that is easy for a human to identify but difficult for a machine. We generally speak of the texture of sand, speckled tile, etc. The simplest texture is of course when all the pixel values are the same. This may be the case for a still lake or a cloudless sky.

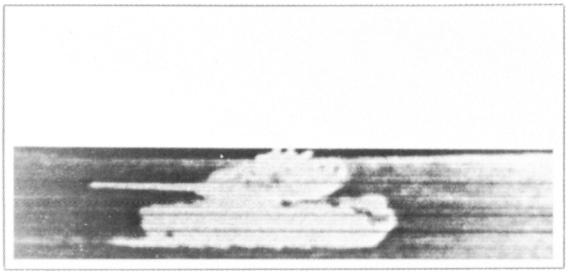


FIGURE 18 The original image.

As an example, consider Figs. 18–21. Figure 18 is the original image (P). The image is 256×256 pixels with each pixel capable of 256 gray levels. Figure 19 is the transformed image [B in Eq. (17) with $\alpha = 0.9$]. Note the scene is still interpretable. Figure 20 has resulted from the opposition's jamming a high-interest portion of the transmission imagery. We denote the jammed NPT by B_0. We note that the sky appears to be cloudless and therefore have a subimage that has a simple texture. If we do not know the exact pixel values, we can assume a statistical average, a constant texture for the sky. The concept is to reconstruct the image P from the jammed image B_0 with a knowledge that a constant texture image segment exists in the original image P.

The solution of the problem is obtained by iteratively computing at the ith stage:

$$\begin{aligned} P_i &= A^{-1} \hat{B}_{i-1} A^{-1} \\ B_i &= A \hat{P}_i A, \end{aligned} \tag{81}$$

where $\hat{B}_i$ and $\hat{P}_i$ are obtained from B_i and P_i by substituting the known values from B_0 and P. That is, the nonjammed values in B_0 will be substituted at each stage

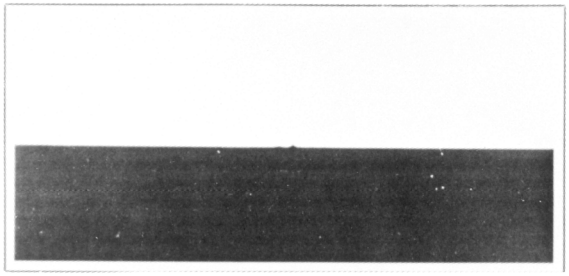


FIGURE 19 The NPT of the original image.

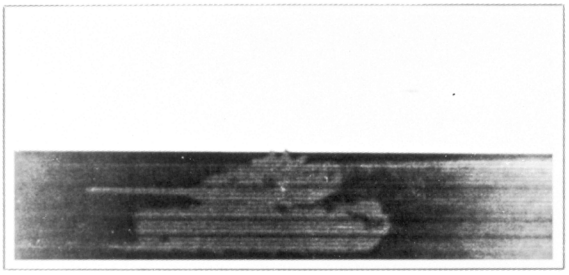


FIGURE 20 The jammed NPT.

in B_i to get B_i and the known values in P (for example, the pixel values corresponding to the constant texture image segment) are substituted in P_i to get $\hat{P}_i$. The algorithm converges and the result after 50 iterations is shown in Fig. 21, which clearly indicates the reconstruction of the jammed image portion.

A few comments are in order as far as the computational complexities. First, the jammed subimage in Fig. 20 consists of 57×256 pixels. If a more common, direct approach were used, such as least squares, we would need to solve $256 \times 57 = 14{,}592$ unknowns in 32,768 equations corresponding to sky pixels. Obviously, the direct approach is computationally intensive. The proposed iterative method involves the computation of Hadamard transforms of vectors of dimension 256. It has been shown that by using a powerful array processor, the proposed method is easily implementable and can operate near real time.

V. THOUGHTS ON THE FUTURE

It is presumptuous, at best, to imagine that one can predict, with any certainty, the future of such an encompassing discipline as signal processing. It seems safe, however, to expect continuing advances in almost all of the areas with, perhaps, the greatest gains in the computational aspects.

Looking back over the history of automatic data processing, we see that hardware dominated the cost of computation in the 1950s and early 1960s. In the latter 1960s, the 1970s, and up through the early 1980s, the dominant cost of computation was software. But, it now seems probable that the next high-cost item will be new computer architectures, their control and calculational flows.

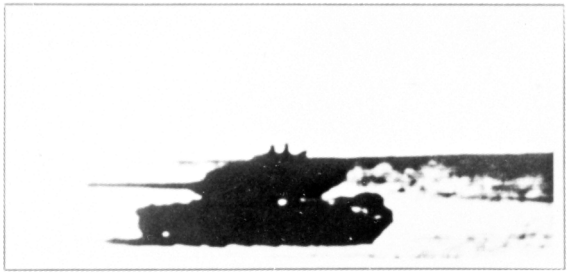


FIGURE 21 Restoration by 50 iterations.

As we look through the history of signal-processing algorithms in the 1970s and 1980s, we come to suspect that the truly phenomenal reductions in work realized by, say, the fast Fourier transform, and other algorithms in those halcyon early days, will probably not be repeated on anything near as significant a scale. What will be breathtaking to watch, however, will be the immense computational power and attendant signal-processing capability available through the intelligent lashing together and control and management of parallel computational structures employing thousands, and eventually millions, of individual processing nodes.

Computers developed around what has become known as the VonNeuman architecture. This simple architecture featured I/O units, a memory unit, a control unit, and an arithmetic and logic unit (ALU). The so-called "bottleneck" in the design arose not from the above modularization of functions but from the practice of passing all of the I/O operations through the ALU. This, of course, halted any computations until the I/O operations were finished. This architecture and data flow were preeminent until 1958.

In the mid-1950s, the Direct Memory Access (DMA) concept began to emerge which freed up the ALU but the controller was still held responsible for I/O control and this still kept the ALU from operating at top, theoretical efficiency. This situation was not remedied until an independent I/O channel architecture was developed and commonly integrated into computer design.

The idea of using more than one processor concurrently developed over a relatively long time period, that is, it is not at all a new idea. As early as 1958, S. H. Unger proposed the concept that computer architecture can be designed to better fit some problems:

> A general purpose digital computer can, in principle, solve any well-defined problem. . . . However, they are relatively inept at solving many problems where the data is arranged naturally in a spatial form. . . . It appears that efficient handling of problems of the type mentioned above cannot be accomplished without some form of parallel action.

Unger's seminal concept has seen realizations from the early work done by D. L. Slotnick and others with the SOLOMON computer to today's COSMIC CUBE at the California Institute of Technology.

Some order, in the form of a hierarchical model of computer organizations, was brought about by M. J. Flynn in 1972. He made popular the terms that have stuck to the

present time and have been adopted worldwide. The terms are best known by their acronyms and are as follows:

1. SISD: *s*ingle-*i*nstruction stream—*s*ingle *d*ata stream: this is the simplest architecture—a single processor.
2. SIMD: *s*ingle *i*nstruction stream—*m*ultiple *d*ata stream: there are three basic types of computer under this rubric.
   a. Array (of) processors: typically an array of processors has one control unit and many independent and identical processors. The processors have their own local memory and are each simultaneously driven by the same instruction from the control unit.
   b. Pipelined processors: execution of operations by time interleaving the component parts of the operation among subprocessor elements tailored to the operation components.
   c. Associative processors: a type of array processor in which the processing elements are not directly addressed but rather impressed into service as needed.
3. MIMD: *M*ultiple-*i*nstruction stream—*M*ultiple *d*ata stream: this is essentially an interacting set of more than one processor. The processors are driven by different programs and they may share the same memory. They may be either "loosely" coupled, "tightly" coupled, or something in between.

The lure of multicomputer processing can be appreciated by looking at an example of what can be done with more than one CPU. Our first such example, for motivational purposes, is the conceptually simple task of adding together a list of numbers.

Let us assume, for convenience, that we have $N = 2^n$ numbers to add together, that is,

$$s = x_1 + x_2 + \cdots + x_{N-1} + x_N. \tag{82}$$

If we use a single processor, we will, of course, have to make $N - 1 = 2^n - 1$ additions. If the time required for a single processor to perform an addition of two numbers is T_A, then the single processor will take $(2^n - 1)T_A$ time units. The single processor will be kept busy for the entire computational time, that is, there will be no idle time in which computational power will be unused. If, on the other hand, we use $N/2 = 2^{n-1}$ processors, we can add all the numbers in $\log_2 2^n = n$ time units. How this is done is shown in Fig. 22. By using 2^{n-1} processors we have achieved a speed advantage of $(2^n - 1)/n$ over a single processor. There are a few points to make:

1. The speed advantage is $\simeq 2^n/n$ which is less than 2^{n-1} for moderately sized n. The time for the parallel computation is *not* reduced in direct proportion to the number of processors.
2. When multiple processors are used on this problem, there is much idle processor time, for example, at addition layer #n only one of the 2^{n-1} processors is busy.

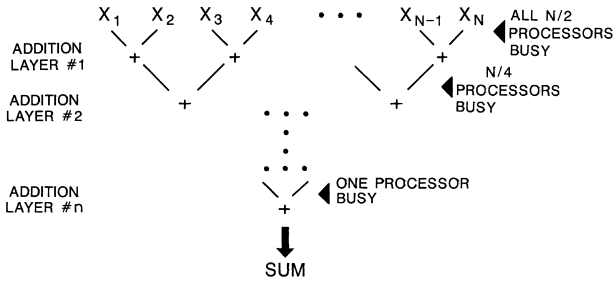


FIGURE 22 Parallel computation of the sum of $N = 2^n$ numbers.

Being a bit more general, if we use $N_p = 2^k$ processors where $\max(N_p) = 2^{n-1}$, we can show that the time to perform the sum s of Eq. (82) is

$$T_s = 2^n / N_p + \log_2 (N_p - 1) \tag{83}$$

and the total processor idle time, which is the sum of the number of processors idle at each step time in forming the sum, is

$$T_i = N_p[\log_2 (N_p - 1)] + 1. \tag{84}$$

We graph T_s and T_i in Fig. 23 for the case for which $n = 10$ and $N = 2^{10} = 1024$.

How well parallel computation works—how much time it is capable of saving—depends almost exclusively on the algorithm used and its implementation. Ware has concocted a model that is very helpful in understanding and appreciating this concept. The "speedup," S, which obtains by using P processors in parallel is defined as

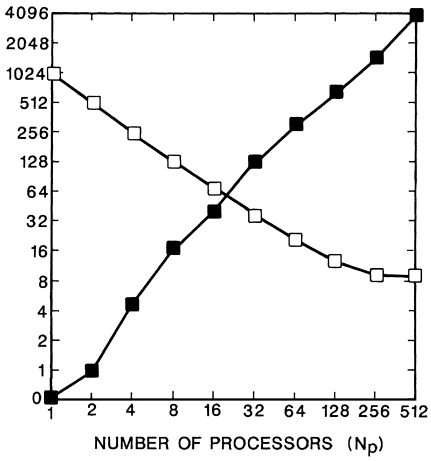


FIGURE 23 T_s (□) and T_i (■) for $N = 2^{10} = 1024$.

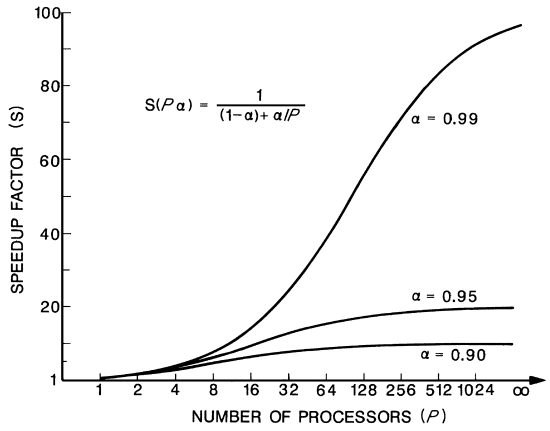


FIGURE 24 S versus P for Ware's model of parallel computation.

$$S = \frac{\text{problem execution time for one processor}}{\text{problem execution time for } P \text{ processors}}. \quad (85)$$

Ware's model introduces the parameter α which is defined as the proportion of work that can be done in parallel. Ware's model assumes that at any time instant all P processors are busy or only one processor is busy. Normalizing the problem execution time for one processor to unity we have

$$S = \frac{1}{(1-\alpha)+\alpha/P}. \quad (86)$$

In Fig. 24 we graph S against P for significant values of α. Note that α must be very close to unity to get anywhere near the full contribution of all P processors. Indeed, as $P \to \infty$, $S \to 1/(1-\alpha)$. Buzbee introduced an additional term into Ware's model. This new term, $\sigma(P)$, reflects the increased "overhead" of managing parallel algorithms such as synchronization and interprocessor communications. With this new term, we have the following expression for S in the "Ware–Buzbee" model for parallel computation:

$$S = \frac{1}{(1-\alpha)+\alpha/P+\sigma(P)}. \quad (87)$$

Now, even if $\alpha = 1$, $\max(S)$ is upperbounded by $1/[1/P+\sigma(P)]$.

As an example of an important process that can be paralleled in a number of ways and as an example of the importance of interprocessor communication times, let us consider the distributed computation of the radix-2 fast Fourier transform (FFT) as developed by L. Bhuyan and D. Agrawal. Consider that we wish to take the FFT of an N time-sampled waveform. For convenience, assume that N is a power of two, $N = 2^n$. Let us assume that the basic unit of FFT computation, the "butterfly," takes B time units to perform. The time to perform the $N = 2^n$ point FFT on a single processor, T_{sp}, is then

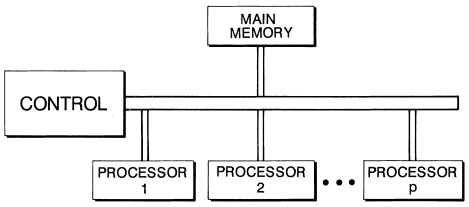


FIGURE 25 Multiprocessors sharing a bus.

$$T_{sp} = nNB/2. \quad (88)$$

Let us now allow $P = 2^m$ processors to work on the FFT in parallel and achieve interprocessor communications by transferring data from processor to processor through a main memory as shown in Fig. 25. Let us assume that a processor-to-processor transfer through main memory requires 2τ time units. Then the time to perform the FFT on the architecture of Fig. 25 is

$$T_{pmm} = (nNB/2P) + mN\tau. \quad (89)$$

The speedup, S, for this configuration is then

$$S = \frac{T_{sp}}{T_{pmm}} = 1 \Big/ \left(\frac{1}{P} + \frac{2m\tau}{nB}\right). \quad (90)$$

Looking at Eq. (90) in terms of the Ware–Buzbee model we see that

$$\alpha = 1 \quad (91a)$$

and

$$\sigma(P) = (2\tau/nB)\log_2 P. \quad (91b)$$

It is thus clear that the dimensionless τ/B is critical to the effective parallelism achieved. We are thus beginning to see that interprocessor communication is an extremely important issue. This leads to the consideration of other architectures such as that shown in Fig. 26. In this schema, the data are not transferred from processor to processor through a main memory but rather directly through an interconnecting network. Interconnecting networks continue to be an extremely important topic for multiprocessor research and they bring their own complications. By necessity, an interconnecting network must, generally, be

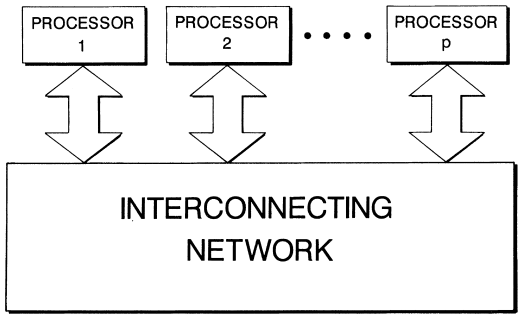


FIGURE 26 Multiprocessors with interconnection network.

capable of executing some impressive combinatorial feats. One commonly used interconnection network is known as the "shuffle-exchange" network. It is a natural for data transfer for the FFT as it is intimately related to the butterfly structure.

However, the shuffle-exchange network takes some time and the network itself requires physical area in the realization medium. All of these tradeoffs must be recognized and appropriately weighted before settling on the final design of a signal processor.

We have caught only a glimpse of the future possibilities afforded by parallel processing. Not even all of the questions are known, and those that are, are not all yet properly phrased. But still we can sense the power that lies ahead and it promises to be as exciting a future as were the "good old days."

SEE ALSO THE FOLLOWING ARTICLES

DIGITAL SPEECH PROCESSING • OPTICAL INFORMATION PROCESSING • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL

BIBLIOGRAPHY

Bhuyan, L., and Agrawal, D. (1982). Applications of SIMD computers in signal processing. *Nat. Comput. Conf*., *AFIPS Conf. Proc*., *Houston*, 135–142.

Byrd, R. H., and Payne, D. A. (1979). Convergence of the Iteratively Reweighted Least Squares Algorithm. Tech. Rep. 313, The John Hopkins Univ., Silver Spring, Maryland.

Flynn, M. J. (1972). Some computer organizations and their effectiveness. *IEEE Trans. Comp*. **C-21**(9), 948–960.

Hall, E. L. (1979). "Computer Image Processing," Academic Press, New York.

Huber, P. J., and Dutter, R. (1974). Numerical solution of robust regression problems. *In* COMSAT 1974, "Proc. Symp. Computational Statistics" (G. Breechmann, ed.), Physica Verlag, Wien, Germany.

Lee, S. H. (ed.) (1981). "Optical Information Processing," Springer-Verlag, Berlin and New York.

Oppenheim, A. V. (1978). "Applications of Digital Signal Processing," Prentice-Hall, Englewood Cliffs, New Jersey.

Oppenheim, A. V., and Schafer, R. W. (1975). "Digital Signal Processing," Prentice-Hall, Englewood Cliffs, New Jersey.

Papoulis, A. (1977). "Signal Analysis," McGraw-Hill, New York.

Rabiner, L. R. (1975). "Theory and Application of Digital Signal Processing," Prentice-Hall, Englewood Cliffs, New Jersey.

Robinson, E. A., and Treitel, S. (1980). "Geophysical Signal Analysis," Prentice-Hall, Englewood Cliffs, New Jersey.

Slotnick, D. L., Borck, W. C., and McReynolds, R. C. (1962). The SOLOMON computer. *AFIPS Conf. Proc*. **22,** 97–107.

Special Issue on Seismic Signal Processing. (1984). *Proc. IEEE* **72.**

Stark, J. (1982). "Applications of Optical Fourier Transforms," Academic Press, New York.

Unger, S. H. (1958). A computer oriented toward spatial problems. *Proc. I. R. E.* (*USA*) **46,** 1744–1750.

Yarlagadda, R., and Hershey, J. (1985). A naturalness preserving transform for image coding and reconstruction. *IEEE Trans. ASSP* **ASSP-33,** 1005–1012.

Yarlagadda, R., Bee Bednar, J., and Watt, T. (1985). Fast algorithms for L_p deconvolution. *IEEE Trans. ASSP* **ASSP-33,** 174–182.



Source Coding, Theory and Applications

N. Farvardin
University of Maryland

I. Introduction
II. Coding Theorems: Theoretical Limits
III. Practical Source Coding Schemes

GLOSSARY

Channel Medium through which information is transmitted.
Channel capacity Quantity, associated with a channel, that determines the largest number of bits that can be transmitted through the channel reliably.
Quantization Mapping (one-dimensional or multidimensional) from a continuous alphabet to a finite alphabet.
Rate–distortion function Function $R(D)$, associated with a source, that represents the smallest amount of information necessary to represent the source with an average distortion less than or equal to D.
Source Information that is to be transmitted to the destination.
Source alphabet Set of possible source output values.

DATA COMPRESSION is the branch of communications that deals with the problem of representing a source of information with as few bits as possible while preserving a certain degree of fidelity. The foundations of data compression were laid in Shannon's original work on source coding with a fidelity criterion in which he established the relationship between the rate-distortion function of the source and the capacity of the channel in a point-to-point communication problem. Data compression schemes, also referred to as source coders or waveform coders, are now being used in various information transmission (storage) situations in which bandwidth (memory) is limited.

I. INTRODUCTION

The ever-growing demand for transmission and storage of data necessitates the more efficient use of existing transmission and storage facilities. A source coding (data compression) system is scheme that operates on source data to remove redundancies so that only those values essential to reproduction are retained. For a given source of information, the optimum performance theoretically attainable by any source coding scheme is given by the source rate-distortion function. Specifically, for a given source, the rate–distortion function $R(D)$ determines the smallest

amount of information necessary to represent the source with an average distortion less than or equal to D.

In any practical situation, given the source probabilistic description, one would try to design the source coding scheme in such a way as to obtain performance as close to the source rate–distortion function as possible. Indeed, practical considerations impose certain serious limitations on the system complexity. In any case, as in other real-world problems, there exists a trade-off between complexity and performance.

In this article, we provide an overall picture of the developments on the theoretical and practical aspects of source coding. To the extent possible, we avoid mathematical details. In particular, the main theorems are stated without proof. The proofs of the theorems can be found in various text books including a few cited in the bibliography at the end of this manuscript. Particular emphasis is placed on providing motivation and intuitive feeling for the different definitions and theorems described herein. Moreover, we describe a few of the more popular practical source coding schemes and provide a performance comparison against the optimum performance theoretically attainable that is described by the source rate–distortion function.

II. CODING THEOREMS: THEORETICAL LIMITS

In this section, we provide a precise formulation of the data compression problem and the basic definitions necessary for subsequent developments. This is followed by the statement of several coding theorems that help shed light on the operational significance of the source entropy, the channel capacity, and the source rate-distortion function.

We start with a discussion of noiseless source coding and the operational significance of the source entropy (or entropy rate) and then extend these results to source coding with a fidelity criterion and fundamental performance limitations dictated by the source rate–distortion function. While most of the results stated herein are for discrete-alphabet sources, at the end of this section extensions to continuous-alphabet sources are also considered.

A. Source Description

Let us assume that the information source is modeled as a discrete-time discrete-alphabet random process $\{U_n\}$ taking values from the alphabet $A_U = \{a_1, a_2, \ldots, a_A\}$. Here, n is simply a time index. Let us assume for the present time that $\{U_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with probabilities $P_i \stackrel{\Delta}{=} P_r(U = a_i)$, $i = 1, 2, \ldots, A$. This source is commonly referred to as a discrete memoryless source (DMS).

The self-information associated with the even $\{U = a_i\}$, $i = 1, 2, \ldots, A$, is defined by the following:

$$I(a_i) = -\log_2 P_i \qquad \text{bits/source symbol.} \tag{1}$$

The quantity $I(a_i)$ is clearly a random variable whose expected value, given by the following:

$$H(U) = -\sum_{i=1}^{A} P_i \log_2 P_i \qquad \text{bits/source symbol,} \tag{2}$$

which is known as the entropy of the DMS.

The self-information associated with an event is inversely proportional to the probability of occurrence of that event. Thus, less likely events contain more self-information and vice versa. In a sense, the self-information of an event can be interpreted as the amount of "surprise" associated with the occurrence of that event. The entropy of a discrete memoryless source, which is the average value of self-information, can be interpreted as the amount of uncertainty associated with the source or, in other words, the entropy of the source is the amount of information that needs to be provided to remove the uncertainty in the source. For data compression of DMSs the source entropy has a very important operational significance. Specifically, the entropy of the DMS determines the smallest number of bits necessary for representing the source with an arbitrarily small probability of error. This fact is more precisely described through the following theorem.

Theorem 1 (Noiseless Source Coding Theorem for the DMS). Consider a discrete memoryless source U with finite entropy $H(U)$. Consider a binary block code that takes sequences of length N of the source output and maps them into blocks of length L of binary digits such that only one source sequence can be mapped into each code sequence. Let P_e, denote the probability of occurrence of a source sequence for which there is no code sequence. Then for any $\varepsilon > 0$, there is a block code such that if

$$\frac{L}{N} \geq H(U) + \varepsilon \tag{3}$$

P_e can be made arbitrarily small by making N sufficiently large. Conversely, if

$$\frac{L}{N} \leq H(U) - \varepsilon \tag{4}$$

there is no code for which P_e can be made arbitrarily close to zero.

This theorem underscores the operational significance of the entropy of the DMS. Note that in Eqs. (3) and (4) the quantity L/N is the number of binary digits (bits)

per source symbol necessary for representing those source sequences for which a code sequence (code word) exists.

Roughly speaking, the theorem establishes that the entropy of a DMS is a fundamental lower limit on the number of bits per source symbol necessary for exact representation of the source. The limit has an important practical implication because it provides us with a means of measuring the performance of a particular noiseless source coding scheme against the very best possible performance.

EXAMPLE 1. Consider a memoryless binary source U with alphabet $A_U = \{0, 1\}$. Let $p = Pr(U = 0)$. Then the entropy of this source $H(U)$ is given by the binary entropy function as follows:

$$H_{b(p)} \triangleq -p \log_2 p - (1-p)\log_2(1-p) \quad \text{bits/source symbol.} \tag{5}$$

Now consider two cases. First, let $p = 0.5$. In this case the source is called a binary symmetric source. The source entropy assumes its largest possible value equal to $H(U) = 1$, bit/source symbol, which implies that the smallest number of bits necessary to represent this source is one bit per symbol. Second, consider the case where $p = 0.1$. In this case the entropy of the source is $H(U) = H_b(0.1) = 0.469$, bits/source symbol. That is, the smallest number of bits necessary to represent this source is even smaller than half a bit per source symbol. Of course, in this case a nontrivial coding operation is needed to represent this source with an average number of bits close to the source entropy. This is interesting because the size of the source alphabet is the same in both cases.

The preceding example shows that knowledge of the probability distribution of the source plays an important role in determining the smallest number of bits necessary for representing the source. This fact can be used advantageously in many communication or storage situations in which the symbols that the source generates are highly different in probability.

Before we extend our results to more complicated sources, let us consider another version of the noiseless source coding theorem that also has some practical implications. The main idea in the following theorem is that of variable-length coding. Here, we assign a binary code word a every source symbol (or sequence of source symbols), but the code words do not necessarily possess the same length. The code words are chosen so that source outputs with large probabilities are mapped to code words of larger lengths in an effort to minimize the average codeword length. In using variable-length codes one has to be careful that the code can be decoded uniquely. That is, for a received sequence of code symbols there is only one decoded sequence. Variable-length codes satisfying these conditions are said to be uniquely decodable.

Theorem 2 (Variable-length Noiseless Source Coding Theorem for the DMS). Consider a DMS U with finite entropy $H(U)$. Suppose a variable-length source coding scheme takes sequences of length N of the source output symbol, say $\mathbf{u}$, and maps them into a binary code word of length $l_N(\mathbf{u})$ for all possible sequences $\mathbf{u}$. Then, there exists a uniquely decodable variable-length code for which the average code word length described by the following:

$$\bar{l}_N = \sum_{\mathbf{u}} Pr(\mathbf{u}) l_N(\mathbf{u}) \tag{6}$$

satisfies

$$\frac{\bar{l}_N}{N} N \leq H(U) + \frac{1}{N}. \tag{7}$$

On the other hand, no such code exists for which

$$\frac{\bar{l}_N}{N} < H(U). \tag{8}$$

This theorem establishes the existence of a uniquely decodable variable-length code for which the average code length per source symbol can be made arbitrarily close to the source entropy by making N large. It is important to mention that there is a major distinction between this theorem and Theorem 2 in that in Theorem 1 there is always a subset of source output sequences for which no code word exists, while in Theorem 2 there is a code word for every possible source output sequence. Therefore, the block coding of Theorem 1 always implies a vanishingly small but nonzero probability of error, while the variable-length code of Theorem 2 does not introduce any coding error.

EXAMPLE 2. Let us consider a DMS U with alphabet $A_U = \{a_1, a_2, a_3, a_4\}$ and probabilities $P_1 = 1/2$, $P_2 = 1/4$, $P_3 = 1/8$, and $P_4 = 1/8$. The entropy of this source is easily calculated to be $H(U) = 1.75$, bits/source symbol. Consider the variable-length code that maps a_1 to 0, a_2 to 10, a_3 to 110, and a_4 to 111. First, note that this variable-length code is uniquely decodable. This is because no code word is a prefix of any other code word. Because of this condition, during the decoding operation, the end of the code words can be easily marked. Variable-length codes that satisfy this condition are called prefix condition codes. For this example the average code word length is given by $\bar{l}_1 = 1.75$ bits/source sample, which is exactly equal to the entropy of the source in this case. Therefore, as one could see from this example, variable-length codes can be used to provide very efficient source coding.

In the past 30 years various algorithms have been developed for the construction of variable-length codes. Among the most important is the Huffman algorithm. For a discrete memoryless source with given letter probabilities, this algorithm constructs a variable-length code that is optimum in the sense that it minimizes the average code word length.

So far we have considered only memoryless sources. Now we extend our results to discrete stationary sources. This class of sources, which includes stationary sources with memory, is representative of many real-world sources and therefore is more important from a practical point of view.

Consider a discrete stationary source $\{U_n\}$ and let $\mathbf{U}_N$ be the generic notation for a sequence of N consecutive source output letters, say, $(U_1, U_2, \ldots, U_N)$. Let $H(\mathbf{U}_N)$ denote the entropy of a source whose output is the N-vector $\mathbf{U}_N$. Then the entropy rate of the discrete stationary source is defined as follows:

$$H_\infty(U) = \lim_{N\to\infty} \frac{1}{N} H(\mathbf{U}_N). \tag{9}$$

The entropy rate of a discrete stationary ergodic source has the same operational significance as the entropy of a discrete memoryless source. Specifically, there are coding theorems that establish that, for a discrete stationary ergodic source with entropy rate $H_\infty(U)$, there exists a code by means of which the source can be represented with arbitrarily small probability of error so long as the average number of bits per source sample is greater than $H_\infty(U)$. Conversely, no such code exists if the average number of bits per source symbol is below the source entropy rate.

In general, the computation of the entropy rate of an arbitrary discrete stationary source is a difficult task. However, for some special sources, the entropy rate can be computed quite easily. An important class of sources that is often used in modeling real-world processes is the class of Markov sources. Such sources are characterized by a set of states, say, $A_S = \{1, 2, \ldots, J\}$, and an output alphabet as before, say, $A_U = \{a_1, a_2, \ldots, a_A\}$. The probability of entering a new state depends only on the most recent state. We denote by S_n the state of the process at time instant n. Let us assume that the conditional probability of entering state i, given that the most recent state was j, is denoted by $\alpha_{ji} \triangleq Pr(S_n = i \mid S_{n-1} = j)$ and that the probability of being in state i is denoted by $\pi_i \triangleq Pr(S_n = i)$ assuming that the Markov chain is irreducible and has reached the steady state. Then the following relationship is used to compute the state probabilities:

$$\pi_j = \sum_{i=1}^{J} \pi_i \alpha_{ij} \qquad j = 1, 2, \ldots, J. \tag{10}$$

Upon defining the conditional entropy of the source, given a particular state i, by the following:

$$H(U \mid i) = -\sum_{k=1}^{A} P_i(a_k) \log_2 P_i(a_k)$$
$$i = 1, 2, \ldots, J, \tag{11}$$

where $P_i(a_k)$ is the probability that the source produces the output a_k when the state is i, it is easy to show that the source entropy rate is given by the following:

$$H_\infty(U) = \sum_{i=1}^{J} \pi_i H(U \mid i). \tag{12}$$

Equation (12) can be used to compute the entropy rate of a Markov source based on the conditional entropies of each state and the corresponding state probabilities.

Now that we have characterized the source and defined the source entropy rate, we describe the communication channel and the channel capacity. From a communication theoretic point of view the interaction between the source and channel and the resulting performance limitations are determined through the source entropy rate and the channel capacity. We elaborate more on this later.

B. Channel Description

We consider only discrete memoryless channels. A discrete memoryless channel (DMC) is a channel with an input alphabet $A_X = \{b_1, b_2, \ldots, b_I\}$ and an output alphabet $A_Y = \{c_1, c_2, \ldots, c_J\}$. At time instant n, the channel maps the input variable X_n into the output variable Y_n in a random fashion. The channel is uniquely determined by specifying the conditional probabilities as follows:

$$q(j \mid i) \triangleq Pr(Y_n = c_j \mid X_n = b_i)$$
$$i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J.$$

Since the channel is memoryless, the conditional probability of a sequence of channel outputs, given a sequence of channel inputs, can be decomposed into the product of the corresponding conditional marginal probabilities, that is:

$$Pr(y_1, y_2, \ldots, y_n \mid x_1, x_2, \ldots, x_n) \prod_{i=1}^{n} Pr(y_i \mid x_i). \tag{13}$$

Definition 1. Consider two discrete random variables X and Y with alphabets $A_X = \{b_1, b_2, \ldots, b_I\}$ and $A_Y = \{c_1, c_2, \ldots, c_J\}$, respectively. Then the average mutual information between X and Y is defined by the following,

$$I(X;Y) = \sum_{i=1}^{I} \sum_{j=1}^{J} P_i q(j \mid i) \log_2 \frac{q(j \mid i)}{Q_j}, \tag{14}$$

where

$$P_i = Pr(X = b_i) \qquad i = 1, 2, \ldots, I$$
$$Q_j = Pr(Y = c_j) \qquad j = 1, 2, \ldots, J$$

and

$$q(j \mid i) = Pr(Y = c_j \mid X = b_i)$$
$$i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J.$$

The average mutual information $I(X;Y)$ is a measure of the amount of "information" that the random variables X and Y provide about one another. Notice from Definition 1 that when X and Y are statistically independent, we have $I(X;Y) = 0$, which means that X and Y do not provide any information about one another. On the other hand, when $X = Y$, we have $I(X;Y) = H(X)$. That is, the amount of information that a random variable X provides about itself is equal to its entropy $H(X)$. This is consistent with out previous interpretation of entropy as a measure of uncertainty in the source.

Definition 2. The capacity of a DMC is defined as follows:

$$C = \max I(X;Y) \qquad \text{bits/channel use,} \qquad (15)$$

where the maximization is taken with respect to all possible input distributions $\{P_i, i = 1, 2, \ldots, I\}$.

The operational significance of the channel capacity resides in the channel coding theorem stated below.

Theorem 3 (Channel Coding Theorem). Consider a discrete memoryless channel with capacity C bits/channel use and a discrete stationary source with entropy rate $H_\infty(U)$ bits/source symbol. Let us assume that the source generates one symbol every τ_s sec and that the channel accepts one channel symbol every τ_c sec. Then, if $H_\infty(U) \le (\tau_s/\tau_c)C$, the output of the source cannot be transmitted over the channel with arbitrarily small probability of error.

Note that the quantity $(\tau_s/\tau_c)C$ is nothing but the channel capacity measured in bits/source symbol. From now on, we refer to this quantity as $\tilde{C}$. Theorem 3 plays a fundamental role in communication theory. It establishes the operational significance of the channel capacity as the rate of transmission below which reliable communication is possible and above which reliable communication is impossible. It is important to note that the theorem establishes the existence of a code that satisfies these conditions and does not provide any insight on the construction of such codes. The construction of implementable codes that guarantee reliable communication remains an open problem.

C. Source Coding with a Fidelity Criterion

While the channel coding theorem establishes the existence of a code that achieves reliable communication when $H_\infty(U) \le \tilde{C}$, it provides no information as to the performance of the best codes when $H_\infty(U) > \tilde{C}$. All it says is that arbitrarily small error probability cannot be achieved. This question was first raised by Shannon 1949 and later solved by himself in 1959. The answer to this question lies in a branch of information theory called rate–distortion theory.

Rate–distortion theory enables us to determine the smallest amount of distortion that needs to be tolerated if indeed the source entropy rate $H_\infty(U)$ is larger than the channel capacity $\tilde{C}$. To be able to define the rate-distortion function of a source, we first have to define a distortion measure.

Let us consider a discrete source with alphabet $A_U = \{a_1, a_2, \ldots, a_A\}$ and a reproduction alphabet denoted by $A_V = \{d_1, d_2, \ldots, d_B\}$. Throughout this discussion we assume that the distortion between the source output u and its reproduction v is measured by a nonnegative, real-valued function $d(u, v)$, $u \in A_U$, $v \in A_V$. This distortion measure is called a single-letter distortion measure because it does not depend on the past or future values of the source or the reproduction sequence. In what follows we provide a definition of the rate–distortion function for a DMS.

Definition 3. Let a DMS with alphabet A_U and probabilities $P_i, i = 1, 2, \ldots, A$, be given and consider a single-letter distortion measure $d(u, v) \ge 0$. Consider a conditional probability assignment $q(j \mid i) = Pr(V = d_j \mid U = a_i)$. This probability assignment results in an average mutual information between the source U and the reproduction V, denoted by $I(\mathbf{q}) \equiv I(U;V)$ and described by the following:

$$I(\mathbf{q}) \equiv I(U;V) = \sum_{i=1}^{A}\sum_{j=1}^{B} P_i q(j \mid i) \log_2 \frac{q(j \mid i)}{\sum_{k=1}^{A} P_k q(j \mid k)} \qquad (16)$$

and an average distortion given by

$$D(\mathbf{q}) = \sum_{i=1}^{A}\sum_{j=1}^{B} P_i q(j \mid i) d(a_i, d_j). \qquad (17)$$

Then the source rate–distortion function is defined as follows:

$$R(D) = \min I(\mathbf{q}), \qquad (18)$$

where the minimization is taken with respect to the set of all conditional probabilities $\mathbf{q} = \{q(j \mid i), i = 1, 2, \ldots, A,\ j = 1, 2, \ldots, B\}$ that satisfy the following:

$$D(\mathbf{q}) \leq D. \tag{19}$$

To describe the operational significance of the rate–distortion function of the source, we have to state a coding theorem. There are various versions of the source coding theorem subject to a fidelity criterion. Let us consider the block coding theorem because it is easier to understand. First, we describe the operation of a block code.

Consider a set $\mathbf{C}$, called the code book, consisting of M vectors $\mathbf{C} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$, each vector consisting of N elements in the reproduction alphabet. For each sequence of N consecutive source outputs, say $\mathbf{u} = (u_1, u_2, \ldots, u_N)$, the encoder selects that code word $\mathbf{v}; \in \mathbf{C}$ that results in the smallest distortion $d(\mathbf{u}; \mathbf{C})$ given by the following:

$$d(\mathbf{u}; \mathbf{C}) = \min_{\mathbf{v} \in C} d(\mathbf{u}, \mathbf{v}), \tag{20}$$

where $d(\mathbf{u}, \mathbf{v})$ is used to denote the distortion between the vectors $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_N)$, described by the following:

$$d(\mathbf{u}, \mathbf{v}) = \sum_{n=1}^{N} d(u_n, v_n). \tag{21}$$

Therefore, the average per-letter distortion associated with this block code is given by the following:

$$D(\mathbf{C}) = \frac{1}{N} \sum_{\mathbf{u}} Pr(\mathbf{u}) d(\mathbf{u}; \mathbf{C}), \tag{22}$$

where $Pr(\mathbf{u})$ is the N-fold probability of the source, and the summation is taken over all possible sequences of length N at the source output. Furthermore, the rate of the code is defined by the following:

$$R = \frac{1}{N} \log_2 M. \tag{23}$$

The average per-letter distortion and the rate associated with a block code are the two important characteristics of the code. The following theorem determines the operational significance of the rate–distortion function through a block coding argument.

Theorem 4 (Source Coding Theorem with a Fidelity Criterion for the DMS). Consider a DMS and a bounded distortion measure $0 \leq d(u, v) \leq d_{\max} < \infty$. Let us suppose that the source rate–distortion function is given by $R(.)$. Then, for a given D and for any $\varepsilon > 0$, there exists a block code $\mathbf{C}$ of rate $R \leq R(D) + \varepsilon$ such that $D(\mathbf{C}) \leq D + \varepsilon$. Conversely, it is impossible to achieve an average distortion less than or equal to D when the code rate R is smaller than $R(D)$.

The important practical consequence of Theorem 4 can be described as follows. If we are willing to accept an average distortion equal to D, then the smallest number of bits necessary for representing the source is given by $R(D)$. In general, for $D > 0$, clearly $R(D) < H(U)$. That is, if we are willing to tolerate some nonzero distortion, we should be able to represent the source with a number of bits less than the source entropy. This is the essence of data compression with a fidelity criterion.

Note that the rate–distortion function is only an attribute of the source and has nothing to do with the channel. To further characterize the rate–distortion function, let us define $D_{\min}$ as the minimum possible value that the average distortion can assume. Also, define $D_{\max}$ as the smallest value of the average distortion for which $R(D) = 0$. Then it is possible to show that for all $D \in (D_{\min}, D_{\max})$, the rate–distortion function is a continuous, strictly decreasing convex $\cup$ function. A typical rate–distortion function curve is depicted in Fig. 1. It should be noted that when the distortion measure is such that small average distortion corresponds to small values of the probability of error, at zero distortion (i.e., the intersection of the $R(\cdot)$ curve with the ordinate) the value of the rate–distortion function is the source entropy, the smallest number of bits necessary to represent the source with arbitrarily small error probability.

Now we can go back to the question raised above as to what happens if we have a source whose entropy rate is bigger than the capacity of the channel. From the channel coding theorem it is clear that to be able to transmit reliably over a channel of capacity $\tilde{C}$, the rate R should be no bigger than $\tilde{C}$. Therefore, if the entropy of the source $H(U)$ is larger than the channel capacity, we can use a source encoder (ideal) to represent the source at a rate just below the capacity at the cost of some distortion given by

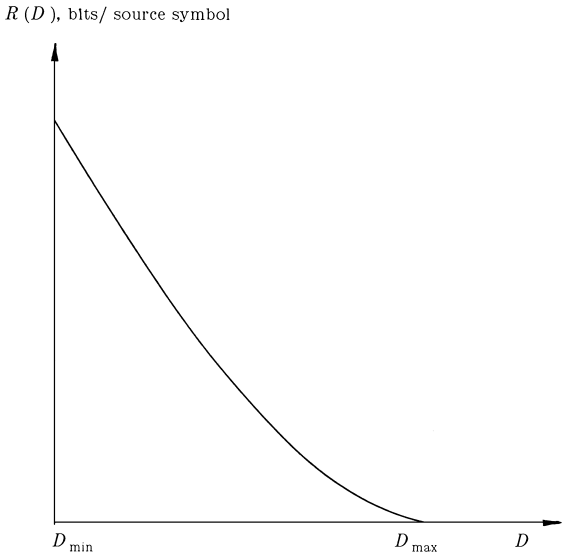


FIGURE 1 Typical rate–distortion function curve.

the relationship $R(D) = \tilde{C}$. Now, the output of the source encoder can be considered as a new source whose entropy is below the channel capacity; hence it can be transmitted reliably over the channel. Therefore, the only source of distortion is that introduced by the source encoder. This observation is more rigorously summarized in the following theorem.

Theorem 5 (Combined Source and Channel Coding Theorem for the DMS and DMC). Consider a DMS with rate–distortion function $R(D)$, bits/source symbol, and a DMC with capacity C, bits/channel use. The source generates one symbol every τ_s sec, and the channel accepts one channel symbol every τ_c sec. Then, for every $\delta\{0\}$, the source can be transmitted over the channel with an average distortion $D + \varepsilon$ where D is such that $R(D) < (\tau_s/\tau_c)(C - \varepsilon)$. Conversely, it is impossible to reproduce the source with an average distortion D if $R(D) > (\tau_s/\tau_c)C$.

EXAMPLE 4. Consider a binary memoryless source U with alphabet $A_U = \{0, 1\}$ which outputs one symbol every $\tau_s = 1$ sec. The source symbol probabilities are $Pr(X = 1) = 1 - Pr(X = 0) = \alpha$. The output of this source is to be transmitted over a binary symmetric channel with identical input and output alphabets $A_X = A_Y = \{0, 1\}$ and crossover probability $Pr(Y = 0 \mid X = 1) = Pr(Y = 1 \mid X = 0) = \beta$. The channel can transmit one channel symbol every $\tau_c = 0.5$ sec. We use the Hamming distance as our distortion criterion. Specifically, we assume that the reproduction alphabet is $A_V\{0, 1\}$ and that $d(u, v) = 0$ if $u = v$, and $d(u, v) = 1$ if $u \neq v$. The rate–distortion function of the source and the capacity of the channel are easily shown to be equal to the following:

$$R(D) = H_b(\alpha) - H_b(D) \qquad \text{bits/source symbol} \tag{24a}$$

and

$$C = 1 - H_b(\beta) \qquad \text{bits/channel use} \tag{24b}$$

respectively. Now, let us assume that the source is symmetric (i.e., $\alpha = 0.5$). Then the first important question is whether it is possible transmit this source over this channel with arbitrarily small probability of error. The answer is yes only if the source entropy is less than the channel capacity. The source entropy, which is nothing but the value of the rate–distortion function at zero distortion, is equal to the following:

$$H(U) = H_b(0.5) - H_b(0) = 1 \qquad \text{bit/source sample.} \tag{24c}$$

Thus, if $C > (\tau_c/\tau_s)H(U)$ or equivalently, $H_b(\beta) < 0.5$, reliable communication is possible. This corresponds to $\beta \lesssim 0.11$. The other question, which is very important, the value of smallest distortion we should tolerate if in fact $\beta > 0.11$. This question is easily answered by equating the rate–distortion function to the channel capacity to get $H_U(D) = 2H_b(\beta)$ 1. For instance, if $\beta = 0.2$, the smallest distortion to be tolerated is $D \cong 0.092$.

We are now ready to extend the above results to more general sources, namely stationary ergodic sources. Indeed, most real-world sources are best modeled by such sources. In general, a typical real-world source is a source that has memory, or in other words, its outputs are correlated. In such cases the advantage of source coding is much more noticeable. This is because, due to statistical dependence between the source outputs, the source can be reproduced with fewer bits than the corresponding memoryless source.

The study of source coding with a fidelity criterion for stationary ergodic sources is more complicated than that of the DMS. The first important issue is that of defining the rate–distortion function for such sources. In what follows we define the rate–distortion function of these sources.

Let us consider a discrete stationary ergodic source $\{U_n\}$ with alphabet A_U, a reproduction alphabet $_V$, and a bounded distortion measure $d(u, v) \leq d_{\max} < \infty$. Let $\mathbf{u} = (u_1, u_2, \ldots, u_N)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ be N-vectors with elements in A_U and A_V, respectively. Then, consider a conditional transition probability assignment $q(\mathbf{v} \mid \mathbf{u})$. This conditional probability assignment results in an average mutual information between the source and the reproduction N-vectors denoted by the following:

$$I(\mathbf{q}) \triangleq I(\mathbf{U}; \mathbf{V}) = \sum_{\mathbf{u}} \sum_{\mathbf{v}} Pr(\mathbf{u}) q(\mathbf{v} \mid \mathbf{u}) \log_2 \frac{q(\mathbf{v} \mid \mathbf{u})}{Q(\mathbf{v})}, \tag{25}$$

where $Q(\mathbf{v})$ is simply the reproduction probability induced by the transition probability $q(\mathbf{v} \mid \mathbf{u})$. Furthermore, an average distortion can be defined between the random N-vectors $\mathbf{U}$ and $\mathbf{V}$ by the following:

$$D(\mathbf{q}) = \sum_{\mathbf{u}} \sum_{\mathbf{v}} Pr(\mathbf{u}) q(\mathbf{v} \mid \mathbf{u}) d(\mathbf{u}, \mathbf{v}). \tag{26}$$

Definition 4. The rate–distortion function of the stationary ergodic source described above is given by the following:

$$R(D) = \lim_{N \to \infty} \frac{1}{N} R_N(D), \tag{27a}$$

where $R_N(D)$ is defined by

$$R_N(D) = \min I(\mathbf{q}) \tag{27b}$$

and the minimization is taken over all possible transition probability assignments that satisfy the following:

$$\frac{1}{N} D(\mathbf{q}) \leq D. \tag{27c}$$

Without repeating the source coding theorem for stationary ergodic sources, we mention here that the rate–distortion function of a stationary ergodic source has exactly the same operational significance as that of the DMS. Various source coding theorems exist to substantiate this.

At this point, we have covered all fundamental results that are relevant to source coding. Before we discuss some specific source coding schemes, there are several issues that deserve more careful consideration.

D. Continuous Alphabet Sources

So far, we have considered only sources that possess a discrete alphabet for both noiseless source coding and source coding with a fidelity criterion. A large class of real-world sources are best modeled by continuous alphabet sources, that is, sources whose set of possible outputs is a continuum of values. Examples are speech and image signals. In what follows, we examine how the concepts of entropy, entropy rate, and rate–distortion function can be extended to sources with continuous alphabets.

First, consider a real-valued discrete-time continuous-alphabet memoryless source U, with alphabet $A_U \subset \mathbb{R}$. Let $f_U(\cdot)$ denote the probability density function of the random variable U. Then, if we consider an interval $(a, a+\Delta] \subset A_U$, its self-information is given by $-\log_2, Pr(a < U \leq a+\Delta)$, which approaches infinity as the length of the interval approaches zero. Roughly speaking, the amount of "surprise" associated with an event $[U = a]$ is infinity. Indeed, if we proceed and define the entropy of the source as the limit when Δ approaches zero of the average value of the above self-information term, we get infinity. This, in fact, makes sense. As we mentioned above, the entropy of the source is a measure of uncertainty or randomness in the source; for a continuous alphabet source it seems intuitively obvious that the source is "infinitely" random. In fact, for a continuous-alphabet source it seems reasonable that an infinitely large number of bits would be needed for an exact representation of the source.

Thus, for a continuous-alphabet source the entropy is not a well-defined quantity. Nevertheless, there exists a similar expression for continuous-alphabet sources, called the source *differential entropy*, defined by the following:

$$h(U) = -\int_{-\infty}^{\infty} f_U(u) \log_2 f_U(u)\, du. \tag{28}$$

While the differential entropy of the continuous-alphabet source carries over some of the properties of the entropy of the discrete source, there are some fundamental differences. First, the differential entropy of a source cannot be considered as a measure of uncertainty of the source. Second, the differential entropy is not necessarily finite, nor is it necessarily nonnegative. Third, in contrast with the entropy of a discrete source, the differential entropy is not invariant under coordinate transformation.

In light of the above discussion, the number of bits for an exact representation of continuous-alphabet sources is infinity. However, it is conceivable that when some average distortion can be allowed, the source can be represented by a finite number of bits per sample. Indeed, the rate–distortion function for a continuous-alphabet source can be defined in a manner similar to that for a discrete source and possesses the same operational significance. Specifically, for the above continuous-alphabet memoryless source, the rate–distortion function is defined as follows:

$$R(D) = \inf I I(q), \tag{29a}$$

where the infinum is taken over the set of all possible conditional probability density functions from the source alphabet A_U to the reproduction alphabet A_V, denoted by $q(v \mid u)$ satisfying the following:

$$D(q) \equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_U(u) q(v \mid u) d(u, v)\, du\, dv \leq D. \tag{29b}$$

Here, $I(q)$ is used to denote the average mutual information between the source and the reproduction variates, defined by the following:

$$I(q) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_U(u) q(v \mid u) \cdot \log_2 \frac{q(v \mid u)}{f_V(v)}\, du\, dv, \tag{30}$$

where $f_V(.)$ is the probability density function of the reproduction random variable V.

The rate–distortion function of the continuous-alphabet memoryless source defined above has an interpretation identical to that of the discrete source described before. Simply, for a given value of D, the corresponding $R(D)$ determines the smallest number of bits necessary to represent the source with an average distortion less than or equal to D. Extensions of the above definition to rate–distortion functions of stationary ergodic sources is essentially the same as that for memoryless sources and is omitted here.

E. Computation and Bounding of the Rate-Distortion Function

The problem of computing the rate–distortion function of a source is a formidable task. The problem is a complex nonlinear constrained optimization problem. Since 1959, when the rate–distortion function of a source was first introduced in Shannon's article, various researchers have been involved in developing techniques for computing this function. To date, there is no known technique for exact

computation of the rate–distortion function of a general source with an arbitrary distortion measure. In fact, what is known is very limited.

Let us consider memoryless sources first. Explicit expressions for the rate–distortion function of such sources is available only for some special cases, including the binary source with the probability-of-error distortion measure and the Gaussian source with the squared-error distortion measure. However, an iterative algorithm, referred to as the Blahut–Arimoto algorithm, exists which can be used for the numerical computation of the rate–distortion function. For sources with memory the problem is much more complex. The only case for which an explicit expression for the rate–distortion function is known is the stationary Gaussian source with the squared-error distortion measure. In this case the rate–distortion function is given in terms of the following parametric expression:

$$D_\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \min(\theta, \Phi(\omega))\, d\omega \tag{31a}$$

$$R(D_\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \max\left(0, \log_2 \frac{\Phi(\omega)}{\theta}\right) d\omega, \tag{31b}$$

where $\Phi(\omega)$ is the source power spectral density.

Due to the fact that for sources with memory the rate–distortion function is defined as a limiting process, extensions of the algorithm for the numerical computation of $R(D)$ *is extremely complex. To date, for sources with memory, no computationally effective algorithm for the numerical computation of* $R(D)$ *is available*.

In view of many difficulties associated with the exact computation of the rate–distortion function, various techniques for developing upper and lower bounds on the rate–distortion function have been developed. In what follows, we describe a few of these bounds that are more commonly used in the literature.

1. Gaussian Upper Bound for Memoryless Sources

Consider a continuous-alphabet memoryless source with a given variance σ^2. For the squared-error distortion measure, the rate–distortion function of this source $R(D)$ is upper bounded by that of a memoryless Gaussian source possessing the same variance, denoted by $R_{\mathrm{G}}(D)$ and described by the following:

$$R(D) \le R_{\mathrm{G}}(D) = \frac{1}{2}\log_2 \frac{\sigma^2}{D}. \tag{32}$$

2. Shannon Lower Bound for Memoryless Sources

Consider a continuous-alphabet memoryless source with a difference distortion measure, that is, $d(u, v) = d(u - v)$. Then the following lower bound on the rate–distortion function holds:

$$R(D) \ge R_{\mathrm{SLB}}(D) = h(U) + \sup_{s\le 0}\left[sD - \log_2 \int_{-\infty}^{\infty} e^{sd(x)}\, dx\right]. \tag{33}$$

In the special case of squared-error, $d(u, v) = (u - v)^2$ the above bound reduces to the following:

$$R(D) \ge R_{\mathrm{SLB}}(D) = h(U) - \frac{1}{2}\log_2(2\pi e D) \quad \text{bits/source symbol.} \tag{34}$$

3. Memoryless Upper Bound for Sources with Memory

Consider a stationary ergodic source with rate–distortion function $R(D)$. Let us denote by $R_1(D)$ the rate–distortion function of a memoryless source with the same marginal density as the original source. Then, the following bound holds:

$$R(D) \le R_1(D). \tag{35}$$

The interpretation of this upper bound is that for a given level of tolerable distortion D, the number of bits necessary to represent a source with memory is smaller than that for the similar source without memory, that is, memory in the source helps represent the source with smaller number of bits.

4. Gaussian Upper Bound for Sources with Memory

Consider a stationary ergodic source with a given power spectral density. Then, for the squared-error distortion measure, the rate–distortion function of this source is upper bounded by that of a Gaussian source with the same power spectral density as described in Eq. (31).

5. Wyner–Ziv Lower Bound for Sources with Memory

For a stationary ergodic source let us define the differential entropy rate of the source by the following:

$$h_\infty(U) = \lim_{n\to\infty} \frac{1}{N} h(U_N), \tag{36}$$

where $h(U_N)$ is the differential entropy of a source whose outputs are N-vectors of the original source. Then the following bound on the rate–distortion function of the original source holds:

$$R(D) \ge R_1(D) - [h(U) - h_\infty(U)]. \tag{37}$$

Note that the quantity $h(U) - H_\infty(U)$ can be interpreted as a measure of memory in the source. In fact, for memoryless sources it is easily shown that $h(U) - h_\infty(U) = 0$.

The above bounds enable us to develop upper and lower bounds on the rate–distortion function of a general source. While, in general, nothing can be said about the tightness of these bounds, in some special cases it is possible to prove that some of the above bounds are tight over some portion of the rate–distortion function.

Now that we are familiar with the fundamental limitations in data compression and the operational significance of the source entropy and rate–distortion function, the important question to be asked is how one should go about designing a source coding scheme whose performance coincides with that of the rate–distortion function of the source. Unfortunately, the coding theorems that establish the operational significance of the rate–distortion function do not have constructive proofs. Therefore, these theorems do not suggest direct design procedures. In general, designing a source coding scheme whose performance coincides with that of the rate–distortion function is out of reach. The best we can hope for is a coding scheme whose performance is close to the rate–distortion function.

In the past 2 decades a tremendous amount of research has been devoted to the development of efficient and easily instrumentable source coding algorithms. In the following section we look at some of the source coding procedures that are used in practical situations and make appropriate performance comparisons.

III. PRACTICAL SOURCE CODING SCHEMES

In the section above we described the basic theoretical limitations in source coding and showed that the source rate–distortion function determines the smallest theoretically achievable rate necessary for representing the source at a given value of distortion. The rate–distortion function of the source provides us with a lower bound on the rate-versus-distortion performance of practical source coding schemes. We now introduce several source coding schemes that are used in practical situations, examine the optimum performance attainable by such schemes, and provide performance comparisons against the corresponding source rate–distortion function.

We begin by zero-memory quantization of memoryless sources because it is the simplest of all schemes considered here and the easiest to understand. We then consider sources with memory and study predictive quantization techniques as well as block transform quantization. We also consider the recently developed algorithm for vector quantization for both memoryless sources and sources with memory.

A. Zero-Memory Quantization of Memoryless Sources

The simplest and most common source coding technique is zero-memory quantization of the real-valued outputs of an appropriately defined source. This technique, often referred to as analog-to-digital conversion (A/D), has been and continues to be a popular method of digitizing analog signals. It is important to emphasize that throughout this article we continue to assume that our signal is a sampled (in time) version of a waveform. If the samples belong to a continuous alphabet, the signal is called an analog signal. On the other hand, if the samples are quantized to a finite number of possible values, the signal is called a digital signal.

Let us assume that the source can be modeled as a discrete-time continuous-alphabet (analog) memoryless source represented by a sequence of (i.i.d.) random variables $[U_n]$ with common p.d.f. $f_U(.)$ possessing a zero mean and variance $\sigma^2 = E[U^2]$.

Consider a zero-memory (or one-dimensional) quantization of the above source. The adjective zero-memory is used to emphasize that the quantizer's output at a given instant depends only on the current input value and is independent of the past and future input and output values.

An N-level zero-memory quantizer is a mapping $q_N(.)$ that maps the source output $u \in \mathbb{R}$ into one of N possible values in the reproduction alphabet $A_V = [y_1, y_2, \ldots, y_N]$. This operation can be described in terms of the thresholds $x_1, x_2, \ldots, x_{N-1}$, partitioning the support of f_U into N disjoint and exhaustive regions according to the following equation:

$$v \triangleq q_N(u) = y_k, x_{k-1} < u \leq x_k, \quad k = 1, 2, \ldots, N, \tag{38}$$

where $x_0 \equiv -\infty$, $x_N \equiv \infty$, and v is the quantizer output.

Here, all x_k are called quantization thresholds, and all y_k are called representation (reconstruction) levels.

An N-level quantizer converts each source output into one of the representation levels and yields an average distortion as follows:

$$D_N(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} f_U(u) d(u, y_k)\, du, \tag{39}$$

where $d(.,\ .)$ is a nonnegative distortion measure and $\mathbf{x}$ and $\mathbf{y}$ are given by $\mathbf{x} = (x_1, x_2, \ldots, x_{N-1})$ and $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, respectively.

Upon defining the following:

$$P_k \triangleq Pr(x_{k-1} < U \leq x_k), \qquad k = 1, 2, \ldots, N, \tag{40}$$

the entropy at the quantizer output is given by the following:

$$H(V) = H_N(\mathbf{x}) = -\sum_{k=1}^{N} P_k \log_2 P_k \quad \text{bits/source symbol.} \quad (41)$$

For a given vector of threshold levels $\mathbf{x}$ and a given vector of representation levels $\mathbf{y}$, the average distortion $D_n(\mathbf{x}, \mathbf{y})$ and the output entropy $H_N(\mathbf{x})$ determine the rate–distortion performance of the quantizer. Here, the quantizer's output entropy $H_N(\mathbf{x})$, in view of the noiseless source coding theorems described in Section II, is representative of the smallest number of bits necessary for representing the quantizer output.

For a given $\mathbf{x}$ and $\mathbf{y}$, we can compare the performance of the quantizer against the rate–distortion function in the following way. Let us consider the rate–distortion plane as illustrated in Fig. 1. Then the pair $(D_N(\mathbf{x}, \mathbf{y}), H_N(x))$ represents a point in this plain. A lower bound to such points for any source coding scheme is given by the source rate–distortion function $R(D)$. Thus the loss in performance introduced by zero-memory quantization is equal to $\Delta R \stackrel{\Delta}{=} H_N(\mathbf{x}) - R(D_N(\mathbf{x}, \mathbf{y}))$, bits per source symbol. Similarly, the loss measured in terms of additional distortion is given by $\Delta D \stackrel{\Delta}{=} D_N(\mathbf{x}, \mathbf{y}) - D(H_N(\mathbf{x}))$, where $D(R)$ is the inverse function of $R(D)$ and is called the source distortion–rate function.

Obviously, one should try to minimize the performance loss given by ΔR or ΔD. In what follows we consider the problem of choosing the vectors $\mathbf{x}$ and $\mathbf{y}$ in an effort to minimize the performance loss.

Definition 5. An N-level quantizer $q_N(.)$ is said to be an optimum entropy-constrained quantizer if it minimizes the average distortion while the output entropy is held below a given value.

According to this definition, the optimum entropy-constrained quantizer minimizes the loss ΔD for a fixed output entropy. More formally, the problem is to minimize the following:

$$D_N(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} f_U(u) d(u, y_k)\, du, \quad (42a)$$

which is subject to the following:

$$H_N(\mathbf{x}) \leq R. \quad (42b)$$

This problem is a nonlinear constrained optimization problem. Necessary conditions for the solution to this problem can be obtained from the celebrated Kuhn–Tucker theorem. Upon letting $(\mathbf{x}^*, \mathbf{y}^*)$ be a relative solution of Eq. (42), there must exist a scalar $\lambda \geq 0$, such that

$$\lambda \ln\left(\frac{P_{k+1}^*}{P_k^*}\right) = d\left(x_k^*, y_{k+1}^*\right) - d\left(x_k^*, y_k^*\right) \quad k = 1, 2, \ldots, N-1, \quad (43)$$

and

$$\lambda\left[H_N(\mathbf{x}^*) - R\right] = 0, \quad (44)$$

where

$$P_k^* = \int_{x_{k-1}^*}^{x_k^*} f_U(u)\, du \quad (45a)$$

and all y_k^* must satisfy

$$\frac{\partial}{\partial y_\kappa} D_N(\mathbf{x}^*, \mathbf{y}^*) = 0. \quad (45b)$$

The above equations are usually extremely difficult to solve. Let us consider the case of the squared-error distortion measure. We maintain this assumption in the rest of this article unless specifically stated otherwise. In this case it is possible to show that Eq. (43) reduces to the following:

$$\lambda \ln\left(\frac{P_{k+1}^*}{P_k^*}\right) = \left(y_{k+1}^* - y_k^*\right)\left(y_{k+1}^* + y_k^* - 2x_k^*\right) \quad k = 1, 2, \ldots, N-1, \quad (46)$$

where

$$y_k^* = \frac{1}{P_k^*} \int_{x_{k-1}^*}^{x_k^*} u f_U(u)\, du. \quad (47)$$

For this special case, there are various algorithms for solving for the optimum $\mathbf{x}^*$ and $\mathbf{y}^*$. An interesting observation to be made from Eq. (47) is that for the squared-error distortion measure, the optimum representation levels are the center of probability mass of their corresponding quantization intervals.

While the above equations can be solved numerically to determine the rate–distortion performance of an N-level quantizer, it is certainly very useful to be able to come up with an explicit expression for the rate–distortion of these optimum quantizers. In general, even for very simple source distributions, it is very difficult to come up with an explicit expression for the rate–distortion performance of optimum entropy-constrained quantizers. However, for a fairly general class of sources, it is possible to show that at high bit rates and for a large number of quantization levels, a quantizer with uniformly spaced thresholds performs very close to optimal. Furthermore, in this case the average distortion of the quantizer as a function of the output entropy R is given by the following simple expression:

$$\hat{D}(R) = \frac{1}{12} 2^{2(h-R)}, \tag{48}$$

where h is simply the source differential entropy. This result was first reported by Gish and Pierce in 1968.

Equation (48) is particularly useful in making comparisons against the source rate–distortion function. Recall from Section II that the Shannon lower bound on the rate–distortion function is given by the following:

$$R_{\mathrm{SLB}}(D) = h - \tfrac{1}{2}\log_2(2\pi e D). \tag{49}$$

It can be shown that under some mild conditions on the source distribution, the Shannon lower bound is tight in the high-bit-rate region. Therefore, the distortion–rate function of the source can be well approximated at high bit rates by the following:

$$\hat{D}(R) = \frac{1}{2\pi e} 2^{2(h-R)}. \tag{50}$$

Now, comparing Eq. (48) with Eq. (50), it is easy to see that $\hat{D}(R)/D(R) = \pi e/6$. This corresponds to 1.53 dB performance loss. Similarly, it can be shown that for a fixed value of distortion D, $\hat{R}(D) - R(D) = 0.5\log_2(\pi e/6) \cong 0.255$ bits/symbol. Here, $\hat{R}(D)$ is the inverse function of $\hat{D}(R)$ described by Eq. (48). These results are extremely useful because they do not depend on the source distribution. While such nice explicit results do not exist for the low-bit-rate region, experimental evidence shows that for most source densities the gap between the source rate–distortion function and the optimum entropy-constrained quantizer performance narrows in the low-bit-rate region. Indeed, at $R = 0$ the two performances coincide.

These results are fundamentally imporant because they determine the optimal performance theoretically attainable by any zero-memory quantization scheme. However, remember that the rate of these quantizers is described in terms of the quantizer output entropy. To represent the quantizer output by a number of bits close to the entropy, some type of a variable-length coding scheme is needed. Variable-length codes suffer from a few difficulties that limit their practical value. Specifically, the decoding of variable-length codes takes place in a sequential fashion. Therefore, an error in the transmission stage may cause a run of errors in the decoded sequence. Furthermore, since most channels operate on a synchronous basis, for the transmission of variable-length codes some type of buffer between the output of the variable-length encoder and the input of the channel is needed. This buffer may overflow or underflow. The overflow and underflow are undesirable and are to be avoided.

In light of these difficulties with the implementation of entropy-constrained quantizers, we now describe a quantization scheme that transmits the representation levels by fixed-length codes and hence does not suffer from the difficulties of variable-length codes.

Definition 6. An, N-level quantizer $q_N(.)$ is said to be optimum in the Lloyd–Max sense, if it merely minimizes the average distortion for the fixed number of levels N.

The rationale behind this definition is that for a fixed number of levels N (assuming that N is a power of 2), we can always represent the quantization levels by $\log_2 N$ bits. Thus, the rate is not dictated by the entropy of the quantizer output but rather by the number of quantization levels. Therefore, in this case the problem is merely to minimize as follows:

$$D_N(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{N} \int_{x_{k-1}}^{x_k} f_U(u) d(u, y_k)\, du, \tag{51}$$

where the quantizer rate is given by $R = \log_2 N$. Similar necessary conditions for the optimality of this quantizer can be derived. Without much ado over the details, let us mention that for the squared-error distortion criterion, the necessary conditions for optimality are given by the following:

$$x_k^* = \frac{y_{k+1}^* + y_k^*}{2} \tag{52a}$$

and

$$y_k^* = \frac{1}{P_k^*} \int_{x_{k-1}^*}^{x_k^*} u f_U(u)\, du. \tag{52b}$$

These conditions were apparently independently derived by Lloyd and Max, hence the term Lloyd–Max quantizer. For a certain class of sources it is possible to prove that the above conditions are also sufficient for the optimality of the quantizer. Several algorithmic procedures for the numerical solution of Eq. (52) are documented in the literature. The most important of all are two algorithms proposed by Lloyd and Max.

Similar to the previous case, in general, in the low-bit-rate region no explicit result for the rate versus distortion performance of the Lloyd–Max quantizer is available. Numerical results, however, for various sources are well documented. There are also some suboptimal schemes for quantization that have certain implementation-related advantages. For example, uniform quantizers are widely used in practical systems. These are quantizers for which the thresholds and the representation levels are uniformly spaced. Another popular technique for quantizer design is what is known as companding. The idea here is to take the source output u and perform some type of nonlinear mapping, say $c(.)$, upon it. The mapping $c(.)$ is called a

compressor. Then the output of the compressor is quantized by means of a uniform quantizer and operated upon by $c^{-1}(.)$. The mapping $c^{-1}(.)$ is called an expander. In the limiting case of a large number of quantization levels, it is possible to show that the optimum compressor is of the following form:

$$\frac{dc_{\text{opt}}(u)}{du} = K \frac{[f_U(u)]^{1/3}}{\int_0^\infty [f_U(x)]^{1/3}\,dx}, \tag{53}$$

where $K = c(\infty)$. In this case, the average quantization distortion is easily shown to be the following:

$$D_{\text{opt},N} = \frac{2}{3N^2}\left\{\int_0^\infty [f_U(u)]^{1/3}\,du\right\}^3. \tag{54}$$

This is another explicit expression for the distortion of quantizers obtained by companding in the high-bit-rate region. It is also shown that quantizers designed by companding are asymptotically (high bit rate) optimal in the Lloyd–Max sense. Therefore, we can use the asymptotic result in Eq. (54) to develop performance comparisons between the Lloyd–Max quantizer and the source rate–distortion function.

EXAMPLE 5. Consider a zero-mean unit-variance Gaussian source U with the following probability density function:

$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}. \tag{55}$$

For this source, the distortion–rate function is given by $R(R) = 2^{-2R}$. The corresponding high-bit-rate Lloyd–Max performance is $D_{\text{opt},N} = \pi\sqrt{3}/(2N^2)$. Noting that the rate is given by $R = \log_2 N$, it is straightforward to show that the distortion–rate performance of the quantizer is given by the following:

$$\tilde{D}_{\text{opt}}(R) = \frac{\pi\sqrt{3}}{2} 2^{-2R}. \tag{56}$$

Comparing this with the source distortion–rate function, one could easily see that $\tilde{D}_{\text{opt}}(R)/D(R) = \pi\sqrt{3}/2$, or 4.35 dB performance loss. The performance loss is easily computed to be 0.722 bits/symbol, substantially larger than that for the entropy-constrained quantizers.

Recall that the gap between the rate–distortion function and the high rate performance of entropy-constrained quantizers is fixed (0.255 bits/symbol) independent of the source distribution. This does not hold for Lloyd–Max quantizers. In other words, the gap between the rate–distortion function and the Lloyd–Max quantizer performance depends upon the source distribution.

It can be concluded that for memoryless sources, zero-memory quantizers perform very close to the optimal performance theoretically attainable. Indeed, for most sources, the optimum entropy-constrained quantizer performs within 0.255 bits/symbol of the best possible performance for all rates. If the source has memory, however, somehow the inherent correlation between neighboring samples must be used in the data compression operation. This cannot be done with a zero-memory quantizer because this quantizer operates on the source outputs instantaneously. Therefore, for sources with memory, to obtain performance close to the rate–distortion function, we need to use something more sophisticated than simple zero-memory quantization.

B. Predictive Quantization

Now we turn our attention to data compression of sources with memory. In sources with memory, there is correlation between adjacent samples, and hence each sample provides some information about the values of the neighboring samples. In other words, there is some redundancy in the source output samples. A "good" data compression system is that which effectively exploits the knowledge that one source sample provides about the neighboring samples, or, equivalently, effectively removes the redundancy in the source.

We now describe a data compression scheme in which instead of quantizing the source outputs directly, we quantize the difference between the source output and its predicted version. Such schemes, in general, are called predictive quantization schemes. The block diagram of a generic entropy coded predictive quantization scheme is illustrated in Fig. 2. Let us first describe the operation of this predictive quantization scheme.

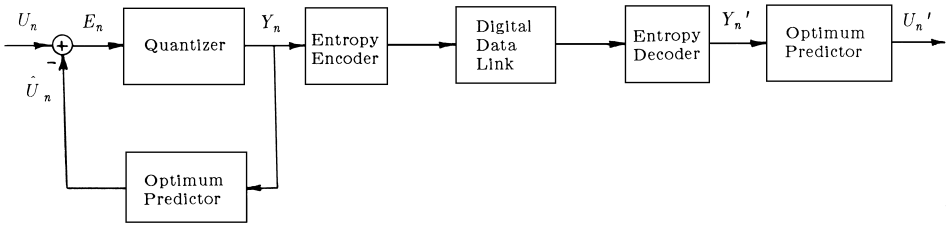


FIGURE 2 Block diagram of a generic predictive quantization scheme.

Roughly speaking, the source signal can be conceived as having two parts: One part ($\hat{U}_n$) is predictable relative to the transmitted sequence Y_n and hence conveys no useful information; the other part (the prediction error, E_n) is unpredictable and (since it uniquely determines the signal) therefore contains the useful information. A predictive quantization scheme attempts to discard the predictable part, for it can be reproduced at the receiver, and encode only the unpredictable portion using a zero-memory quantizer.

Let us assume that the source U_n is a stationary ergodic source with zero-mean and variance σ_U^2. At time instant n, the predicted estimate of U_n based on all the past transmitted values $Y_1, Y_2, \ldots, Y_{n-1}$ is given by the following conditional expectation:

$$\hat{U}_n = E\{U_n \mid Y_{n-1}, Y_{n-2}, \ldots, Y_1\}. \tag{57}$$

On the other hand, the instantaneous estimate of U_n based on observing $Y_1, Y_2, \ldots, Y_n$ is described by the following:

$$\tilde{U}_n = E\{U_n \mid Y_n, Y_{n-1}, \ldots, Y_1\}. \tag{58}$$

At time instant n, the predictor in the transmitter loop makes an estimate of the value of U_n based on the previously transmitted values. Note that, assuming the transmitted sequence Y_n reaches the receiver with no error (i.e., the transmission medium is noiseless), the quantity $\hat{U}_n$ can also be produced at the receiver. Thus the transmitter computes the part of the source output that could not be predicted (the prediction error) as follows:

$$E_n = U_n - \hat{U}_n \tag{59}$$

and quantizes it for transmission.

The analysis of the predictive quantization in such generality as just described is formidable. Even the computation of the average distortion is not a straightforward matter. So, to be able to carry the analysis further we make some specific assumptions on the structure of the source. For simplicity, we assume that the source is a first-order autoregressive process described by the following:

$$U_n = \rho U_{n-1} + W_n \qquad n = 1, 2, \ldots, \tag{60}$$

where W_n is a sequence of i.i.d. random variables with zero mean and variance σ_W^2 and ρ the source correlation coefficient.

With this assumption, it is possible to show that

$$\hat{U}_n = \rho \tilde{U}_n. \tag{61}$$

Furthermore, we can write the following:

$$\tilde{U}_n = \hat{U}_n + E\{E_n \mid Y_n, Y_{n-1}, \ldots, Y_1\} \quad n = 1, 2, \ldots. \tag{62}$$

If we further assume that the sequence E_n is nearly independent, we can rewrite Eq. (62) as the following:

$$\tilde{U}_n = \hat{U}_n + E\{E_n \mid Y_n\} \qquad n = 1, 2, \ldots. \tag{63}$$

Finally, if we assume that the quantization levels are chosen as the center of probability mass of their respective intervals, it is easy to show that

$$E\{E_n \mid Y_n\} = Y_n, \qquad n = 1, 2, \ldots. \tag{64}$$

Therefore, with all the above assumptions, we have the following:

$$\tilde{U}_n = \hat{U}_n + Y_n \tag{65a}$$

and

$$\hat{U}_n = \rho \tilde{U}_n. \tag{65b}$$

The predictive quantization scheme described by Eq. (65) is widely used in practice and is known as the differential pulse code modulation (DPCM). Various versions of the DPCM scheme are being used in different speech and image compression situations.

In what follows we analyze the performance of the predictive coding scheme described above. To be consistent with our previous results we perform our analysis in a rate–distortion theoretic framework.

The average distortion associated with the above DPCM scheme is given by the following:

$$D_N(\mathbf{x}, \mathbf{y}) = E\left\{(U_n - \tilde{U}_n)^2\right\}, \tag{66}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the quantization thresholds and representation levels as before, and error-free transmission is assumed so that $U_n' = \tilde{U}_n$. It is interesting to note that

$$U_n - \tilde{U}_n = E_n - Y_n, \tag{67}$$

which means that to minimize the overall reconstruction error, it suffices to minimize the quantization error. Therefore, the problem of optimizing the performance of the DPCM scheme is reduced to that of designing the quantizer optimally.

While the optimization of the performance of DPCM may appear trivial, it is actually a formidable task. The main difficulty resides in the fact that to design the quantizer, we need to have the p.d.f. of the error sequence. However, the p.d.f. of the error sequence is not known and depends on the structure of the quantizer. So, clearly, there is a very intricate intertwining between the p.d.f. of the prediction error and the parameters of the quantizer. In fact, even for a fixed, maybe nonoptimal quantizer, the computation of the p.d.f. of the prediction error involves the solution of a complicated integral equation. Therefore, in general, it is very difficult to optimize the performance of the DPCM scheme. For first-order Gauss–Markov and

Laplace–Markov sources, however, some numerical results on the performance of optimal DPCM systems are available.

In the high-bit-rate region, as in zero-memory quantization, it is possible to come up with explicit formulas for the rate versus distortion performance of optimum DPCM systems. Specifically, when the source is a first-order Gaussian autoregressive source (i.e., the source described in Eq. (60) with Gaussian W_n, it is possible to show that in the high-bit-rate region the prediction error has a Gaussian density as well, with the following variance:

$$\sigma_E^2 = \sigma_W^2 + \rho^2 D, \tag{68}$$

where D is the average quantization distortion given by the following:

$$D \equiv \hat{D}(R) = \frac{1}{12} 2^{2(h_E - R)} \tag{69}$$

and h_E is the differential entropy of the prediction error sequence. Simplification of these equations implies the following:

$$\hat{D}(R) = \left(\frac{\pi e}{6}\right) \sigma_W^2 \, 2^{-2R}, \tag{70}$$

a result very similar to the asymptotic performance of the optimum zero-memory quantizer at high bit rates. Indeed, this is a nice way of assessing the performance improvement obtained by DPCM over zero-memory quantization. If we had used zero-memory quantization to encode this source, the average distortion at high bit rates would be exactly the same as that in Eq. (70) with σ_U^2 replacing σ_W^2. Noting that

$$\sigma_U^2 = \frac{\sigma_W^2}{1 - \rho^2} \tag{71}$$

we would conclude that the DPCM coding gain is given by the following:

$$G = \frac{1}{1 - \rho^2}. \tag{72}$$

Thus, for $\rho = 0$, there is no coding gain. This is to be expected as $\rho = 0$ corresponds to a memoryless source in this case, and there is nothing to be gained by using a predictive coding scheme. On the other hand, for values of ρ close to 1, corresponding to a highly correlated source, the coding gain is very large.

Another interesting comparison to be made is against the source rate–distortion function. It is possible to show that the source rate–distortion function at high bit rates is given by the following:

$$R(D) = \frac{1}{2} \log_2 \frac{\sigma_W^2}{D}. \tag{73}$$

Comparison of Eq. (73) with Eq. (70) immediately implies that the performance loss of the optimum entropy-constrained DPCM at high bit rates is exactly 0.255 bits/symbol, as in the memoryless case. The results for the case in which a Lloyd–Max quantizer is used in the DPCM loop essentially follow the same pattern as in the zero-memory case and hence is not repeated here.

C. Block Transform Coding

As mentioned above, when there is memory in the source, the data compression system should attempt to make use of the memory. The DPCM scheme utilizes the memory in the source by making a prediction of the current sample based on the previous transmitted samples. An alternative would be to quantize a block (preferably large) of the source outputs at once. This way, we can make use of the correlation in the source to reduce the transmission rate. Indeed, in the proof of the source coding theorem subject to a fidelity criterion, we did resort to a block coding argument. There are two difficulties involved in this type of quantization. First, how to choose the quantization representation vectors and, second, the complexity of encoding. We address both of these issues in Section III.D where we discuss vector quantization. However, in this subsection, we present a particular form of block quantization that is commonly referred to as block transform coding.

For the sake of consistency, let us continue to assume that our source is described by the model in Eq. (60). Let us denote by $\mathbf{U}$ a vector of L consecutive source samples, say, $\mathbf{U} = (U_1, U_2, \ldots, U_L)^T$. In a block transform coding scheme the source output vector $\mathbf{U}$ is operated upon by a nonsingular $L \times L$ transformation matrix $\mathbf{A}$ to generate an L-vector $\mathbf{Y}$ with uncorrelated components described by the following:

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_L)^T = \mathbf{AU}. \tag{74}$$

The intuitive idea behind this transformation is to exploit the correlation between the source samples by first generating a set of uncorrelated variables $Y_1, Y_2, \ldots, Y_L$ and then encoding these sample by sample. Thus each component of $\mathbf{Y}$ is quantized and encoded by means of a zero-memory quantizer. Therefore, we need a total of L quantizers, one for each transform coefficient. These quantizers could be entropy-constrained or Lloyd–Max quantizers. Let us denote the quantized version of Y_k by $\hat{Y}_k$.

In the receiver, an $L \times L$ transformation matrix $\mathbf{B}$ operates upon the received vector $\hat{\mathbf{Y}}$ to generate a replica of $\mathbf{U}$, say $\hat{\mathbf{U}}$, given by the following:

$$\hat{\mathbf{U}} = \mathbf{B}\hat{\mathbf{Y}}. \tag{75}$$

For the squared-error distortion measure, the average distortion per source symbol is given by the following:

$$D = \frac{1}{L} E\{(\mathbf{U} - \hat{\mathbf{U}})^T (\mathbf{U} - \hat{\mathbf{U}})\}. \tag{76}$$

At this point we are ready to describe the problem. Given a fixed number of bits/source symbol, our goal is to minimize the average distortion given by Eq. (76). The minimization of D is achieved by the appropriate choice of the matrices $\mathbf{A}$ and $\mathbf{B}$ and the appropriate design of the quantizers used for encoding the transform coefficients. This problem was first studied by Huang and Schultheiss in 1963; they considered the block transform quantization of stationary Gaussian sources with the assumption that Lloyd–Max quantizers are used for encoding the transform coefficients.

Huang and Schultheiss showed that for a given $\mathbf{A}$ the best matrix $\mathbf{B}$ is the inverse of $\mathbf{A}$. Furthermore, they showed that the best transformation matrix $\mathbf{A}$ is the matrix whose rows are the orthonormalized eigenvectors of the covariance matrix of $\mathbf{U}$. This optimum transformation, called the Karhunen–Loeve transformation, is a unitary transformation (i.e., $\mathbf{A}^T = \mathbf{A}^{-1}$). Thus it is easy to show that

$$\begin{aligned} D &= \frac{1}{L} E\{(\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}})\} \\ &= \frac{1}{L} \sum_{i=1}^{L} \lambda_i D_i, \end{aligned} \tag{77}$$

where λ_i is the variance of the ith transform coefficient and D_i is the variance-normalized distortion associated with the ith quantizer.

Let us assume that the rate associated with the ith quantizer is denoted by R_i. Now that the transformation matrices $\mathbf{A}$ and $\mathbf{B}$ are chosen, the problem is simply to minimize the following:

$$D = \frac{1}{L} \sum_{i=1}^{L} \lambda_i D_i \tag{78a}$$

subject to

$$\frac{1}{L} \sum_{i=1}^{L} R_i = R. \tag{78b}$$

Notice that at this point we are still free to choose our quantizers to be entropy-constrained or Lloyd–Max quantizers. If we choose to use entropy-constrained quantizers, R_i, should denote the entropy at output of the ith quantizer. On the other hand, should we decide to use Lloyd–Max quantizers, $R_i = \log_2 N_i$, where N_i is the number of quantization levels for the ith quantizer.

In either case, the problem as stated in Eq. (78) is that of determining the optimum assignment of rates (entropies or bits) to different transform coefficients. Therefore, the problem of optimal design of block transform quantizers is reduced to that of an optimal rate assignment problem. This problem, for the case of Lloyd–Max quantizers, was solved by Huang and Schultheiss. The extension to entropy-constrained quantization has also been considered recently.

For the sake of completeness, we must mention that for Gaussian sources and in the limit of large block sizes, it is established that the block transform quantization has a performance identical to DPCM at high bit rates. This holds for both entropy-constrained and Lloyd–Max quantization. In the low-bit-rate region, explicit expressions for the performance of the block transform quantization are not available. However, numerical results indicate that block transform quantizers have performance superior to the DPCM scheme.

Block transform quantization schemes and various adaptive versions thereof are used in many practical source coding schemes. Many low-bit-rate speech coding and video conferenceing schemes are based on this system. An attribute of transform coding schemes that makes them very attractive is the low complexity of these systems. Once the transformation is performed, the rest is simple, zero-memory quantization. In fact, due to the very structure of the block transform coding scheme, the system is amenable to parallel processing, which is becoming very popular. To minimize the complexity of computing the matrix transformation, various suboptimal transformations have been considered in different applications. The discrete cosine transform has proven to be a very useful transformation in the sense that it provides performance very close to that obtained by the optimal Karhunen–Loeve transformation, and yet it can be implemented by means of very fast algorithms.

D. Vector Quantization

Vector quantization is simply a multidimensional extension of the zero-memory (one-dimensional) quantization scheme. An L-dimensional, N-level vector quantizer $q_N(.)$ is a mapping from $\mathbb{R}^L$ to the set of reproduction vectors $A_Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$. Here, each $\mathbf{y}_i$ is a vector in $\mathbb{R}^L$. Associated with the vector quantizer is a partition of $\mathbb{R}^L$, say $A_S = \{S_1, S_2, \ldots, S_N\}$, where each $S_i \subset \mathbb{R}^L$. Here, the A_S is a collection of disjoint and exhaustive regions in $\mathbb{R}^L$. The operation of the vector quantizer is then described by the following:

$$q_N(\mathbf{u}) = \mathbf{y}_i \qquad \text{if} \quad \mathbf{u} \in S_i. \tag{79}$$

Here, $\mathbf{u} = (u_1, u_2, \ldots, u_L)$ is an L-dimensional vector of source outputs.

This operation is identical to the one we used in the proof of the source coding theorem subject to a fidelity criterion. Thus we know that in the limit of large L, there

exists a vector quantizer with appropriate A_Y and A_S, whose performance is very close to the rate–distortion function of the source.

Ever since Shannon established the usefulness of block source encoding techniques, there has been considerable interest in finding good block encoders or vector quantizers. Until 1980, the developments on this subject were limited to the developments of uniform vector quantizers for which some asymptotic properties were also developed. A breakthrough in the area of design of optimal vector quantizers was made in 1980 by Linde, Buzo, and Gray. They provided an iterative algorithm, referred to hereafter as the LBG algorithm, for the design of optimal error minimizing vector quantizers based on a training sequence of data. Essentially, the LBG algorithm is a multidimensional extension of an algorithm for one-dimensional quantizer design suggested by Lloyd. Since the invention of the LBG algorithm, a tremendous amount of research activity has been expended on various applications and refinements of vector quantization. We briefly describe the LBG algorithm.

Let us suppose that the distortion is given by a nonnegative function $d(\mathbf{u}, \mathbf{v})$. For a fixed number of representation vectors N, we wish to minimize the average distortion as follows:

$$D = \frac{1}{L} E\{d(\mathbf{U}, \mathbf{V})\}, \tag{80}$$

where $\mathbf{V}$ is used to denote the quantizer output vector. Let us suppose that the partition A_S is fixed. Then, it is possible to show that the best set of reproduction vectors A_Y is such that $\mathbf{y}_i$ satisfies

$$E\{d(\mathbf{U}, \mathbf{y}_i) \mid \mathbf{U} \in S_i\} = \min_{\mathbf{x}} E\{d(\mathbf{U}, \mathbf{x}) \mid \mathbf{U} \in S_i\}. \tag{81}$$

Conversely, assume that the set of reproduction vectors A_Y is fixed. Then it is easy to show that the best partition A_S is such that

$$\mathbf{u} \in S_i \quad \text{only if} \quad d(\mathbf{u}, \mathbf{y}_i) \leq d(\mathbf{u}, \mathbf{u}_j) \quad \text{all } j. \tag{82}$$

The LBG algorithm is simply a successive application of the conditions described by Eqs. (81) and (82). In practice, the actual implementation of the algorithm is not based on computing the expectation in Eq. (81). In fact, computing this expectation in high dimensions involves multidimensional integration over regions of possibly irregular shapes and, hence, is computationally very difficult. Linde, Buzo, and Gray, in their original article, suggested a remedy to this problem by using a training sequence of data which, in turn, replaces multidimensional integration, by simple summation.

An additional advantage of this algorithm is that the average distortion decreases at each step of the algorithm, and hence the LBG algorithm guarantees convergence to a locally optimum quantizer.

A few comments about the performance and complexity of the LBG-based vector quantization are in order. First, it should be noted that in the absence of an entropy constraint, for a given block size L, the vector quantizer provides the best possible performance. Of course, there is no guarantee that the locally optimum quantizer that the algorithm has converged to is also globally optimum. The choice of the initial conditions dictates the locally optimum quantizer the algorithm converges to. Unfortunately, there is no way a priori of determining the initial choice that results in the globally optimum solution.

To provide a complete comparison, in Tables I and II we consider the performance of optimum DPCM, optimum

TABLE I Signal-to-Noise Ratios (in Decibels) of Optimum Entropy-Constrained Block Transform Quantization and Other Schemes for a First-Order Gauss–Markov Source with $\rho = 0.5$

	Consecutive source samples (L)					
R (bits/sample)	1	2	4	8	DPCM[a]	$R(D)$[b]
1	4.58	5.43	5.78	5.96	5.22	7.27
	(4.40)[c]	(4.71)	(5.30)	(5.41)		
	[4.40][d]	[5.18]	[5.69]			
2	10.51	11.14	11.45	11.60	11.58	13.29
	(9.30)	(9.33)	(10.12)	(10.24)		
3	16.53	17.15	17.47	17.62	17.77	19.31
	(14.62)	(14.62)	(15.34)	(15.46)		

[a] Optimum differential pulse code modulation.
[b] Rate–distortion function.
[c] Huang–Schultheiss scheme, values in parentheses.
[d] Vector quantization scheme, values in brackets.

TABLE II Signal-to-Noise Ratio (in Decibels) of Optimum Entropy-Constrained Block Transform Quantization and Other Schemes for a First-Order Gauss–Markov Source with $\rho = 0.8$

	Consecutive source samples (L)					
R (bits/sample)	1	2	4	8	DPCM[a]	$R(D)$[b]
1	4.58	7.45	8.47	9.02	7.56	10.46
	(4.40)[c]	(6.87)	(7.90)	(8.39)		
	[4.40][d]	[6.85]	[8.18]			
2	10.51	12.74	13.86	14.41	14.44	16.48
	(9.50)	(11.29)	(12.64)	(13.20)		
3	16.53	18.75	19.84	20.31	20.96	22.50
	(14.62)	(16.93)	(17.78)	(18.44)		

[a] Optimum differential pulse code modulation.
[b] Rate–distortion function.
[c] Huang–Schultheiss scheme, values in parentheses.
[d] Vector quantization scheme, values in brackets.

block transform quantization (both entropy-constrained and Lloyd–Max), and optimum vector quantization, for a first-order Gauss–Markov source as described in Eq. (60) with $\rho = 0.5$ and $\rho = 0.8$, respectively. It should be noted that the results for the vector quantizer are only available at $R = 1$ bit/sample. At higher bit rates the computational requirement of the vector quantizer design is beyond current practical limits. In Tables I and II, the results corresponding to $L = 1$ describe the performance of the zero-memory quantizer. The performance gains of the DPCM, block transform coding, and vector quantization compared with zero-memory quantization are obvious. Notice that these gains are more substantial for the larger value of ρ.

Note that the rate associated with the vector quantizer is given by $R \triangleq (\log_2 N)/L$, or $N = 2^{LR}$. Therefore, the number of representation vectors grows exponentially with the rate R and the block size L. This exponential growth limits both the design and the implementation of vector quantizers at high rates and for large block sizes. In an actual implementation of the vector quantizer, the distortion between the source output vector and each representation vector is computed, and the vector with smallest distortion is used to represent the source vector. Therefore, the number of distortion computations is directly proportional to the number of representation vectors. To overcome these difficulties, various suboptimal schemes for the implementation of the vector quantizers are under investigation.

Vector quantization has found practical applications in many source coding situations. It has made a considerable effect on the performance of speech encoders based on linear predictive coding. It has also had an impact on speech recognition research. More recently, some researchers have studied the performance of vector quantizers in image coding situations.

E. Relevant Issues

Before closing this section, we wish to describe briefly two issues that are inherently important in any data compression situation.

1. Sensitivity to Transmission Errors

The function of a source encoder is to remove redundancy in the source. This removal of redundancy results in additional sensitivity to transmission errors. This is an inherent problem with data compression schemes. However, some data compression schemes, due to their very structure, are more sensitive to this problem. For instance, variable-length codes could be extremely sensitive to channel errors due to the sequential nature of the decoding operation. Currently most data compression schemes are designed based on the assumption that the channel is noiseless. It is felt that if the data compression scheme is designed based on the assumption that the channel is actually noisy, considerable performance improvement could be obtained.

2. Adaptivity and Robustness

In most data compression schemes it is assumed that the source statistical knowledge is available. In practical cases, however, either a complete probabilistic description of the source is not available or the source statistics undergoes temporal and/or spatial variations. There are two ways of combating this problem. First, one could design the source encoder in such a way that it performs well for a class of possible source statistics, in which case the design is called robust. Second, the system could be designed to be adaptive in the sense that it senses variations in the source statistics and adapts itself to the current statistics by changing some parameters. Both of these issues are fundamentally important in the design of any data compression system for real-world applications.

SEE ALSO THE FOLLOWING ARTICLES

DATABASES • DATA STRUCTURES • DATA TRANSMISSION MEDIA • NETWORKS FOR DATA COMMUNICATION • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL • VOICEBAND DATA COMMUNICATIONS

BIBLIOGRAPHY

Berger, T. (1971). "Rate–Distortion Theory: A Mathematical Basis for Data Compression," Prentice-Hall, Englewood Cliffs, NJ.

Davisson, L. D. (1972). Rate-distortion theory and applications. *Proc. IEEE* **60,** 800–808.

Farvardin, N., and Modestino, J. W. (1984). Optimum quantizer performance for a class of non-Gaussian memoryless sources. *IEEE Trans. Inform. Theory* **IT-30,** 485–497.

Farvardin, N., and Modestino, J. W. (1985). Rate–distortion performance of DPCM schemes for autoregressive sources. *IEEE Trans. Inform. Theory* **IT-31,** 402–418.

Gray, R. M., ed. (1982). Special issue on quantization. *IEEE Trans. Inform. Theory* **IT-28,** 127–262.

Gray, R. M. (1990). "Source Coding Theory," Kluwer, Boston.

Hankerson, D. R. (1997). "Introduction to Information Theory and Data Compression," CRC Press, Boca Raton, FL.

Jayant, N. S., and Noll, P. (1984). "Digital Coding of Waveforms—Principles and Applications to Speech and Video," Prentice-Hall, Englewood Cliffs, NJ.

Linde, Y., Buzo, A., and Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Trans. Commun.* **COM-28,** 84–95.

Makhoul, J., Roucos, S., and Gish, H. (1985). Vector quantization in speech coding. *Proc. IEEE* **73,** 1551–1588.

Marcellin, M. W., and Fischer, T. R. (1990). Trellis coded quantization of memoryless and Gauss–Markov sources. *IEEE Trans. Commun.* **38,** 82–93.

Salomon, D. (2000). "Data Compression: Complete Reference," 2nd ed., Springer-Verlag, Berlin.

Sayood, K. (2000). "Introduction to Data Compression," 2nd ed., Kaufmann, Los Altos, CA.

Schuster, A., and Katsaggelos, C. (1997). "Rate–Distortion Based Video Compression: Optimal Video Frame Compression and Object Bound Encoding," Kluwer Academic, Boston.

Yip, P. C., and Rao, K. R. (2000). "Transform and Data Compression Handbook," CRC Press, Boca Raton, FL.



Speech Synthesis Based on Linear Prediction

Bishnu S. Atal
AT&T Bell Laboratories

I. Introduction
II. Basics of Linear Prediction
III. Linear Prediction of Speech Signals
IV. Applications of Linear Prediction

GLOSSARY

A/D converter Device that samples a continuous signal at uniformly spaced time intervals and converts each sample into an integer number.

Autocorrelation function The autocorrelation function of a signal $f(t)$ is given by

$$R(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t) f(t+\tau)\, dt.$$

The autocorrelation function without the limit over T extending to ∞ is called the short-time autocorrelation function of $f(t)$.

Convolution The convolution of two signals $f(n)$ and $g(n)$ is given by

$$x(n) = \sum_{k=1}^{K} f(n-k) g(k).$$

Decibel (dB) Representation of energy on a logarithmic scale. The energy E in decibels is given by

$$E_{\text{db}} = 10 \log_{10} E.$$

Delta function Function that is zero everywhere except at one value of its argument and whose integral over the entire range, for which the function is defined, is equal to unity.

Fourier spectrum The squared magnitude of the Fourier transform $F(\omega)$ is called the Fourier spectrum $G(\omega)$. That is,

$$G(\omega) = |F(\omega)|^2$$

Fourier transform The Fourier transform of a signal $f(t)$ is given by

$$F(\omega) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-j\omega t}\, dt,$$

where $\omega = 2\pi f$ is the frequency f expressed in radians.

Frequency-domain techniques Signal processing techniques in which the signal is analyzed using either Fourier transforms or bandpass filters.

Linear prediction spectrum Spectrum determined by linear prediction analysis method.

Nyquist rate Lowest rate at which a finite-bandwidth signal can be periodically sampled in order to reproduce

the signal completely and faithfully. The Nyquist rate is equal to twice the bandwidth of the signal.

Pitch period Fundamental period of voiced speech. The pitch period varies continuously in speech production.

Power spectrum The power spectrum of a signal $f(t)$ is given by

$$G(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-j\omega\tau} d\tau,$$

where $R(\tau)$ is the autocorrelation function of $f(t)$.

Pulse code modulation (PCM) Method of sending an analog signal on a digital communication channel by quantizing each sample of the signal.

Quantizer Device that converts a real number into an integer selected from a given set.

Unvoiced speech Speech sounds generated by forcing air through a narrow constriction in the vocal tract.

Vocoder Speech transmission system in which a small number of slowly varying signals are sent on the communication channel to synthesize the speech signal at the receiver. The word vocoder is a short form of voice coder.

Voiced speech Speech sounds produced by periodic opening and closing of the vocal cords. Typical examples of voiced sounds are vowels.

White noise Noise whose spectrum is "flat" as a function of frequency. The autocorrelation function of white noise is zero at all nonzero delays τ.

PREDICTION deals with extrapolating the present from the past. The prediction can in general be nonlinear. However, linear prediction is simpler and a fairly complete mathematical theory is available. In linear prediction, each piece of the past is weighted and added together to form an estimate of the present. The prediction is not perfect whenever the present includes some new information. The theory of linear prediction plays an important role in the study of physical phenomena, such as speech signals, that can be represented as a time series (set of observations arranged as a function of time). Linear prediction is one of the most widely used techniques for speech analysis. It provides a simple and effective method of representing speech signals in terms of a small number of slowly varying parameters.

I. INTRODUCTION

Linear prediction has become an important method for speech analysis and synthesis. Although the basic mathematical foundations of the theory of linear prediction were established in the early 1940s, the method did not find widespread use for speech analysis and synthesis until the mid-1960s. Most speech analysis until then was based exclusively on frequency-domain techniques. Since then, the linear prediction has become the predominant method of speech processing. What are the reasons linear prediction has become so important for speech processing?

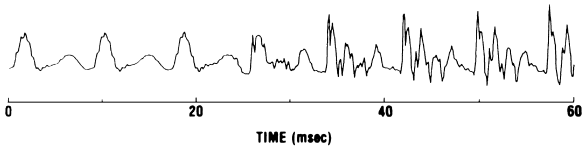


FIGURE 1 Sound pressure for a short segment of speech signal, about 60 ms in duration.

The reasons will become evident if we look briefly at how the speech signals are produced by the human vocal system. There are many different modes in which speech is produced. One important mode—vowel production—relies on the regular and almost periodic opening and closing of the vocal cords to convert the steady air flow from the lungs into a wideband signal rich in harmonics. This wideband signal excites the resonances of the vocal cavities, which are then radiated from the mouth into the open air. The waveform of sound pressure for a short speech segment is illustrated in Fig. 1, where the regular motion of the vocal cords appears as periodic variations in the sound pressure and the vocal resonances appear as oscillations within each period.

Digitized speech waveform, produced by sampling natural speech at 8000 samples/s (0.125 ms time intervals), provides a fairly accurate representation of the original speech signal. In speech analysis, our goal is to represent speech by a small number of slowly varying parameters reflecting the changing nature of the shape of the vocal cavities and the motion of the vocal cords. The theory of linear prediction provides an accurate representation of speech in terms of a few (as low as 12) parameters specified once every 10 ms.

One is often interested in identifying vocal resonances independent of the motion of the vocal cords. It is difficult to determine these resonances in a spectrum obtained by using Fourier transforms or bandpass filters. This is illustrated in Fig. 2, which shows the distribution of energy in different frequency components of the speech waveform of Fig. 1. It is obvious that the harmonic structure in the spectrum introduced by the vocal cords makes it very difficult to identify vocal resonances. The linear prediction methods take account of resonances directly in the analysis procedure and are thus able to provide a smooth spectrum with well-defined peaks corresponding to the resonances. Such a spectrum for the speech waveform of Fig. 1 is shown in Fig. 3.

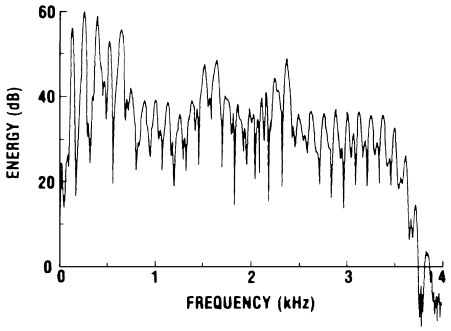


FIGURE 2 Energy in different frequency components of the speech waveform shown in Fig. 1. The abscissa represents the frequency in kilohertz (kHz) and the ordinate represents the energy in decibels (dB).

Consider now an application such as digital speech coding. Here, linear prediction offers an easy solution. The speech waveform is a continuously varying (analog) signal. In speech coding, we are interested in transmitting speech over a digital communication channel that can carry only the binary digits zero and one. How can we convert the analog speech signal into a digital form using as few binary digits (bits) as possible? We can reduce the bit rate needed to transmit the speech signal significantly by sending only new information on the digital channel. If we look at the speech waveform of Fig. 1, we notice that there is considerable similarity between the signal and its preceding values. Now, it would be wasteful to send that information on the digital channel that can be predicted from the past. The theory of linear prediction allows us to determine exactly what is predictable in the signal and remove that information from the speech signal before transmission on the digital channel.

Speech signals can be analyzed using both the frequency-domain approach based on Fourier transforms and the time-domain approach based on linear prediction. Although the frequency- and time-domain approaches appear different from each other, there is a close connection between the two for stationary signals. This connection is carried over only loosely to nonstationary signals, such as speech, whose characteristics vary slowly with time.

We shall discuss in this article the basic concept of linear prediction, the merit of using linear prediction for speech analysis and synthesis, and the differences between linear prediction and more traditional spectral estimation techniques. We will also discuss some important applications of linear prediction in the coding and synthesis of speech signals.

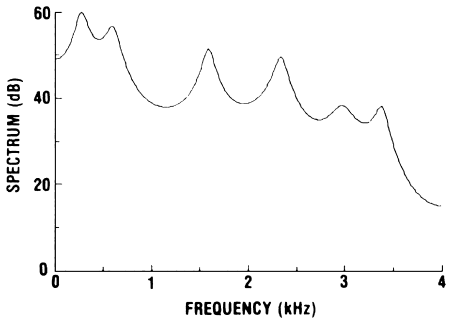


FIGURE 3 Spectrum of the speech waveform of Fig. 1 using linear prediction techniques. The abscissa and ordinate are the same as in Fig. 2.

II. BASICS OF LINEAR PREDICTION

Since the idea of linear prediction is so closely related with the resonances, we will illustrate the basic principle of linear prediction by looking at a sine wave. A sine wave of frequency f_0 has an equation

$$s(t) = a \sin(2\pi f_0 t + \phi), \tag{1}$$

where a is its amplitude and ϕ is its phase. An example of a sine wave is illustrated in Fig. 4a. The spectrum of a sine wave has only one frequency component at the frequency f_0, as shown in Fig. 5. A sine wave can be sampled, and there is no loss of information if the sampling frequency is higher than twice the frequency of the sine wave. Let us assume that we are sampling the sine wave at time intervals 0.125 ms apart, which corresponds to a sampling rate of 8000/s. The sampled version is illustrated in Fig. 4b and is represented mathematically as

$$s(n) = a \sin(2\pi f_0 nT + \phi), \tag{2}$$

where T is the sampling interval. If we look at successive samples, $s(n)$, $s(n-1)$, and $s(n-2)$, of the sine wave, we find an interesting relationship between the three:

$$s(n) = 2\cos(2\pi f_0 T)s(n-1) - s(n-2). \tag{3}$$

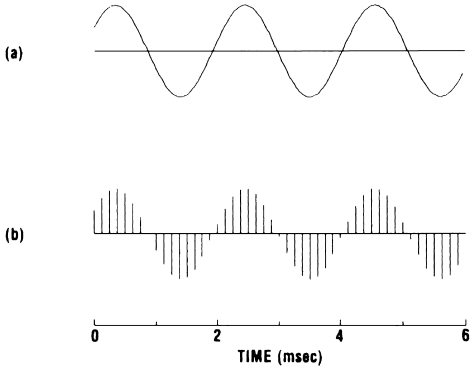


FIGURE 4 (a) Waveform of a sine wave of frequency $f_0 = 475$ Hz. The phase ϕ is $\pi/4$. (b) The same sine wave sampled at 8000 samples/s.

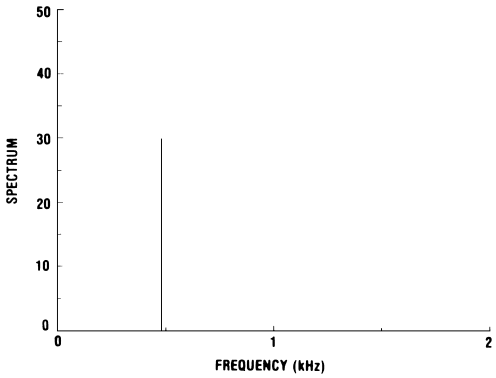


FIGURE 5 Spectrum of infinite duration sine wave of frequency $f_0 = 475$ Hz.

Thus, every sample is exactly determined from the previous two samples, if the frequency of the sine wave is known. Conversely, the frequency of the sine wave can be determined from just three consecutive samples of the waveform. A sine wave is thus a completely predictable signal; that is, every sample can be predicted exactly from the past values.

The spectrum of a sine wave of *infinite* duration has only a single frequency component, as shown in Fig. 5. However, this property is not true for the spectrum of a sine wave of *finite* duration, as shown in Fig. 6. The spectrum of a finite segment of a sine wave is distributed around the frequency f_0 with one major and a number of smaller side lobes. The problem of identifying the frequency components in the spectrum becomes even more difficult for signals containing many sine wave components. This is illustrated in the waveform shown in Fig. 7 consisting of two sinusoidal components with frequencies at 1050 and 1475 Hz, respectively. The Fourier spectrum of the finite duration signal, shown in Fig. 8, produces considerable smearing and a number of side lobes that make the task

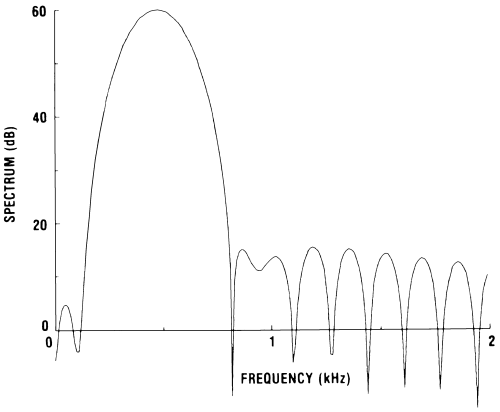


FIGURE 6 Spectrum of a finite segment of sine wave of frequency $f_0 = 475$ Hz.

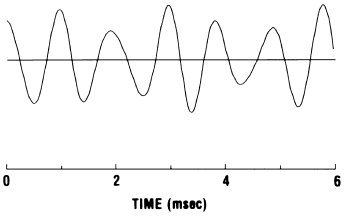


FIGURE 7 Waveform of a signal with two sine wave components at 1050 and 1475 Hz.

of determining the frequencies of sine wave components very difficult.

Most physical signals change their characteristics as a function of time. Thus, we must compute the Fourier spectrum over short segments of time, where the signal is quasi-stationary. Linear prediction analysis provides an alternative route to obtain a frequency spectrum, where the problems associated with spectral smearing and side lobes are avoided. For a composite signal consisting of two sine waves, the linear prediction must include four past samples of the signal, two for each sinusoidal component. Figure 9 shows an example of the spectrum obtained by linear prediction analysis for the waveform shown in Fig. 7. The sine wave components are clearly resolved in the linear prediction spectrum. We will discuss the detailed procedure for computing the spectrum using linear prediction later in this article.

III. LINEAR PREDICTION OF SPEECH SIGNALS

A. Prediction Error

We saw earlier that for a sine wave each sample could be predicted as a linear combination of the previous two

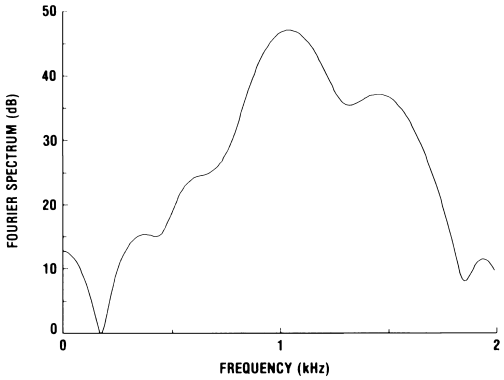


FIGURE 8 Fourier spectrum of the signal shown in Fig. 7. The abscissa shows the frequency in kilohertz, and the ordinate is the energy (in decibels) of a frequency component. The frequency component at 1475 Hz is not resolved in the spectrum.

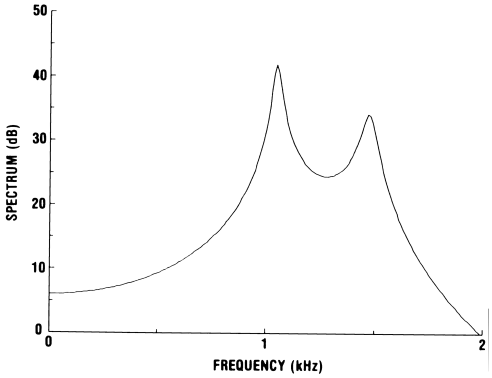


FIGURE 9 Spectrum obtained by linear prediction analysis. The two frequency components at 1050 and 1475 Hz are clearly visible in the spectrum.

samples. We now generalize prediction to include several past samples, as it applies to more complex signals such as speech. In general, prediction is not perfect and some error, usually referred to as *prediction error*, is unavoidable. The linear estimate of the nth speech sample based on past speech samples is given by

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k). \tag{4}$$

We call $\hat{s}(n)$ the predicted value of $s(n)$ and the coefficients a_k the predictor coefficients. We define now the prediction error $\varepsilon(n)$ as the difference between $s(n)$ and $\hat{s}(n)$, that is,

$$\varepsilon(n) = s(n) - \hat{s}(n). \tag{5}$$

So far, we have not said how the predictor coefficients are to be found. It is reasonable to choose them so as to minimize the prediction error. However, it will not be appropriate to minimize the prediction error at each sampling instant n by a new selection of predictor coefficients. Rather, we define a mean-squared prediction error averaged over an interval of N samples, that is,

$$E_p = \frac{1}{N} \sum_{n=1}^{N} \varepsilon^2(n) \tag{6}$$

and choose the predictor coefficients to minimize the prediction error E_p.

As we mentioned earlier, speech is a nonstationary signal and its prediction characteristics vary with time, dependent on the changes in the vocal system. In practical terms, a new set of predictor coefficients could be obtained once every 10 ms, an interval short enough to capture most of the important changes in the speech signal. The averaging interval N has to be large enough to provide a reliable estimate of the error E_p but it should not be too large to smooth the variations in the predictor coefficients. A suitable value of N is 160, corresponding to a time interval of 20 ms for speech sampled at 8000 samples/s.

B. Determination of Predictor Coefficients

Let us discuss now the method of determining the predictor coefficients for a given speech segment. The predictor coefficients that minimize the mean-squared prediction error E_p are obtained by setting the partial derivative of the error with respect to each a_k equal to zero. The error as a function of the predictor coefficients is given by

$$E_p = \frac{1}{N} \sum_{n=1}^{N} \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2. \tag{7}$$

We set the partial derivatives $\partial E_p / \partial a_k$ equal to zero:

$$\frac{\partial E_p}{\partial a_k} = 0, \quad k = 1, 2, \ldots, p. \tag{8}$$

This gives p simultaneous equations in p unknowns:

$$\sum_{k=1}^{p} a_k \phi_{ik} = c_i, \quad k = 1, 2, \ldots, p, \tag{9}$$

where

$$\phi_{ik} = \sum_{n=1}^{N} s(n-i)s(n-k)$$

and

$$c_i = \sum_{n=1}^{N} s(n)s(n-i). \tag{10}$$

Equation (10) can be expressed in matrix notations as

$$\Phi \mathbf{a} = \mathbf{c}, \tag{11}$$

where the element in the ith row and kth column of $p \times p$ matrix Φ is ϕ_{ik}, a_k is the kth element of the p-dimensional vector $\mathbf{a}$, and c_i is the ith element of the p-dimensional vector $\mathbf{c}$. The unknown predictor coefficients can be determined by solving the p simultaneous equations given in Eq. (9) or the matrix equation given in Eq. (11).

We have not yet discussed an important question—what is the correct value of p? As we mentioned earlier, each resonance adds two additional predictor coefficients to the prediction. In speech there are approximately four or five resonances and it would be necessary to set $p = 10$. But we must also add a few extra coefficients for the excitation produced by the vocal cords, the coupling of the nasal tract, and the sound radiation from lips. This suggests that $p = 12$ is a reasonable estimate of the prediction order. Figure 10 shows a plot of the root mean squared (rms) prediction error as a function of prediction order p for both voiced and unvoiced speech. It is clear that not much

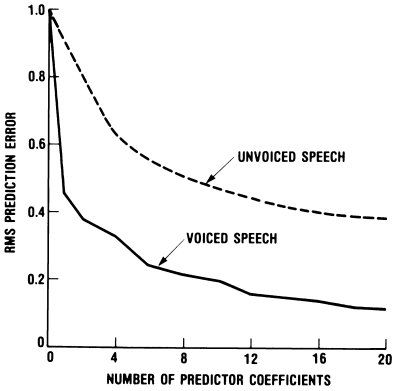


FIGURE 10 Normalized root mean squared (rms) prediction error plotted against p, the prediction order, for voiced and unvoiced speech.

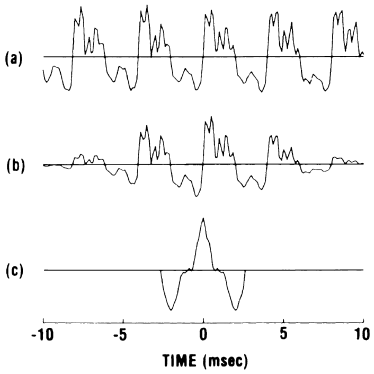


FIGURE 11 Windowing of speech segment for computing the short-time autocorrelation function. The figure shows (a) the speech waveform, (b) the windowed speech waveform, and (c) the autocorrelation function computed by Eq. (10).

has to be gained by selecting a value of p much greater than 12.

C. Autocorrelation Approach

We now consider a special case when the averaging interval in computing the matrix elements ϕ_{ik} is extended to infinity. In this case, the matrix elements ϕ_{ik} depend only on the index $|i-k|$. This just states that the autocorrelation function of an infinite sequence is determined by the relative delay $|i-k|$ between $s(n-i)$ and $s(n-k)$. The $p \times p$ matrix Φ is now a symmetric Toeplitz matrix, where all the elements along a given subdiagonal are identical.

The above approach requires that the averaging interval for computing the mean-squared prediction error be infinite. However, we are interested in minimizing the mean-squared prediction error averaged over N samples. One way to avoid this conflict is to set the speech samples identically equal to zero outside the averaging interval. That is, we set $s(n) = 0$, for $n < 1$ and $n \geq N$, in computing the correlations. However, a sharp truncation of the speech segment is likely to create a large increase in prediction error at the beginning and the end of the segment being analyzed. This problem can be avoided by using tapered windows, whose amplitude falls to zero in a gradual manner.

Equation (9) then simplifies to

$$\sum_{k=1}^{p} c_{|i-k|} a_k = c_i, \quad k = 1, 2, \ldots, p, \tag{12}$$

where the correlation coefficients c_i are similar to the autocorrelation function except that the summation is carried over a finite interval from $n = 1$ to N, as indicated in Eq. (10). We will call these coefficients short-time autocorrelations. To avoid sharp discontinuities in the signal caused by truncation of the waveform at $n = 1$ and $n = N$, the speech signal must be multiplied by a smoothly tapered window rather than a rectangular window. This procedure is illustrated in Fig. 11, where the curve marked (a) shows the original speech waveform, (b) shows the windowed speech waveform, and (c) shows the short-time autocorrelation function. A suitable window for this application is a raised cosine window given by

$$w(n) = 0.5 + 0.5 \cos[2\pi(n-1)/N]$$
$$n = 1, 2, \ldots, N. \tag{13}$$

Due to the Toeplitz nature of the matrix Φ, the solution of Eq. (11) is considerably simplified. An efficient method developed by Durbin and Levinson exists for solving such matrix equations and requires much less computational effort than is needed to solve p simultaneous equations in p unknowns. In this method, the predictor coefficients a_k are obtained recursively for successively increasing values for the order of prediction. The coefficients at the ith stage of recursion are given by

$$a_k^{(i)} = a_k^{(i-1)} - a_i^{(i)} a_{i-k}^{(i-1)},$$
$$1 \leq k \leq i, 2 \leq i \leq p, \tag{14}$$

where

$$a_i^{(i)} = \frac{\left[c_i - \sum_{k=1}^{i-1} c|i-k| a_k^{(i-1)} \right]}{\left[c_0 - \sum_{k=1}^{i-1} c_k a_k^{(i-1)} \right]}, \quad 2 \leq i \leq p. \tag{15}$$

Let us now look at the results of linear prediction analysis of the speech segment shown in Fig. 11 as the order of prediction is increased. Figure 12 shows the waveforms of the original speech signal $s(n)$ and the prediction error $\varepsilon(n)$ for different values of p from 2 to 10. The prediction error does not change much from $p = 8$ to $p = 10$ for this particular segment. The sharp pulses in the prediction

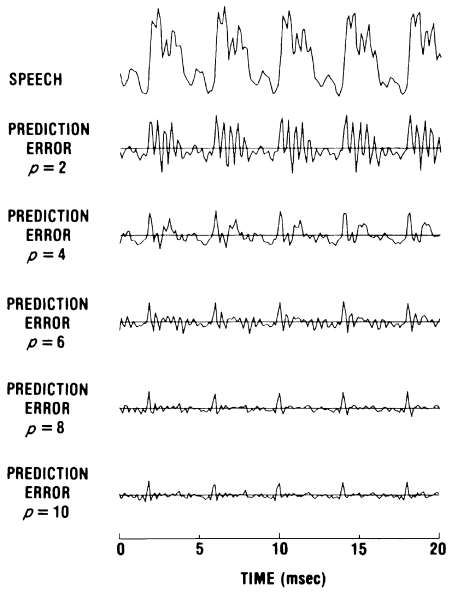


FIGURE 12 Waveforms of speech and prediction error signals for different values of p, the prediction order.

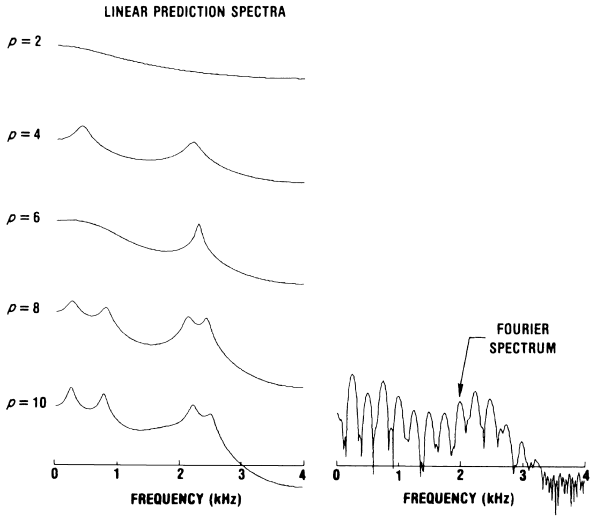


FIGURE 13 Linear prediction spectra computed from the predictor coefficients for different values of p, the prediction order. The Fourier spectrum is also shown on the right side for comparison.

error reflect the sharp closing action of the vocal cords. The excitation change is so rapid at these instants that the waveform cannot be accurately predicted.

We had seen earlier that the predictor coefficients could be used to determine the frequency of a sine wave. We will now generalize this result for more complex signals, such as speech. It is possible to define a power spectrum $G(f)$ from the predictor coefficients. The exact relationship can be expressed as

$$G(f) = \frac{E_p}{\left|1 - \sum_{k=1}^{p} a_k e^{-2\pi j f k l f_s}\right|^2}, \qquad (16)$$

where f_s is the sampling frequency and E_p is the minimum value of the mean-squared prediction error. Figure 13 shows the linear prediction spectra obtained for different values of p from 2 to 10 for the speech segment shown in Fig. 11. The different resonances of the vocal cavities are clearly visible in the linear prediction spectrum. The Fourier spectrum is also shown on the right side in the figure for comparison. The figure again confirms that there is very little change in the spectrum from $p = 8$ to $p = 10$.

D. Partial Correlation Approach

Partial correlations play an important role in linear prediction analysis. The concept of partial correlation is important in determining the *true intrinsic correlation* between two variables. In statistics, a partial correlation refers to the correlation between two variables after the portion in both that can be predicted from a third variable has been eliminated. For speech samples spaced m sample intervals apart, the partial correlation is defined as the correlation between the speech samples after the influence of intervening samples has been removed by linear prediction. Consider a pair of speech samples $s(n)$ and $s(n-m)$. The intervening samples are $s(n-1)$, $s(n-2)$, ..., $s(n-m+1)$ and the partial correlation r_m is defined as

$$r_m = \frac{\sum_{n=1}^{N} \varepsilon(n)\nu(n)}{\left[\sum_{n=1}^{N} \varepsilon^2(n) \sum_{n=1}^{N} \nu^2(n)\right]^{1/2}}, \qquad (17)$$

where $\varepsilon(n)$ and $\nu(n)$ (often called forward and backward prediction errors) are given by

$$\varepsilon(n) = s(n) - \sum_{k=1}^{m-1} a_k s(n-k)$$

and

$$\nu(n) = s(n-m) - \sum_{k=1}^{m-1} b_k s(n-k), \qquad (18)$$

and $a_1\, a_2, \ldots, a_{m-1}$; $b_1, b_2, \ldots, b_{m-1}$ are two sets of predictor coefficients chosen to minimize the mean-squared forward and backward prediction errors, respectively. In principle, the partial correlations can be computed directly from Eq. (17). However, partial correlations can be computed using a recursive solution similar to one used in the autocorrelation approach.

The magnitude of each partial correlation by definition is ≤ 1. Furthermore, if the averaging interval for computing the partial correlations is extended to infinity, it can be

shown that the ith partial correlation is identical to the intermediate coefficient $a_i^{(i)}$ obtained in the autocorrelation approach. Thus, a direct conversion of partial correlations to predictor coefficients is possible by replacing a_k by r_k in Eq. (16). However, it turns out that in many applications there is no need to do this conversion.

E. Pitch Interval Correlations

As we mentioned earlier, there are many different modes in which the speech signals are generated by the human vocal system. Voiced speech, which includes vowels and vowel-like sounds, is produced as a result of modulating the air flow from the lungs by the opening and closing of the vocal cords. Unvoiced speech, which includes certain consonants and fricatives, is produced as a result of air flow from the lungs passing through a narrow constriction in the vocal tract to generate turbulence. Often, both vocal cord vibrations and turbulence are present simultaneously, and we then get mixed-mode sounds. Examples of speech waveforms are shown in Fig. 14, which shows voiced speech, unvoiced speech, and mixed-mode speech. The random nature of unvoiced speech is caused by the turbulence. Notice that the regular opening and closing of the vocal cords produce a repetitive speech pattern with a definite period, called the pitch period. The durations of pitch periods in speech range from 5 to 20 ms for men and from 2.5 to 10 ms for women.

So far, we have discussed procedures that used only a small number of past speech samples to form the predicted value. The repetitive nature of voiced speech can also be exploited in the prediction. In fact, adjacent pitch periods are often very similar, and thus the predicted value based on speech samples in the previous pitch period can be very effective in reducing the prediction error. Let us discuss this kind of prediction, often referred to as pitch prediction, in more detail.

A simple method of including a pitch period in the prediction will be to define a predicted value $\hat{s}_p(n)$ as

$$\hat{s}_p(n) = \beta s(n - M), \tag{19}$$

where M is the duration of pitch period in samples and β is the pitch predictor coefficient. We can determine both of the unknown parameters β and M by using a procedure similar to the one we used before for determining the predictor parameters. We will again choose the predictor parameters to minimize the prediction error. We are going to skip the mathematical details here and provide the final result of this minimization.

Let us look at the autocorrelation function c_i, as defined in Eq. (10). For pitch prediction, the index i in c_i should include the highest expected value of the pitch period, which could be 20 ms. An autocorrelation function typical for vorced speech is shown in Fig. 15 and as expected shows a high correlation at pitch intervals. The best value of M that minimizes the prediction error is the index i at which the correlation c_i is maximum. The best value of β that minimizes the prediction error is given by

$$\beta = \frac{\sum_{n=1}^{N} s(n)s(n - M)}{\sum_{n=1}^{N} s(n - M)s(n - M)}. \tag{20}$$

It might appear from the above discussion that we have to know whether speech is voiced or not to make use of pitch prediction. That is not so. The linear prediction technique is general enough to accommodate all classes of speech—voiced, unvoiced, and mixed-mode. An example of the autocorrelation function for unvoiced speech is illustrated in Fig. 16, and the correlations at large delays

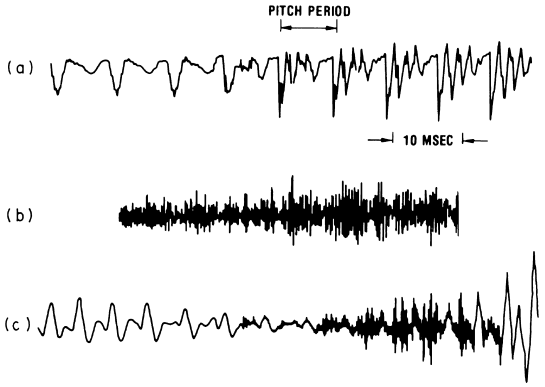


FIGURE 14 Three examples of speech waveforms showing different modes of excitation of the vocal cavities: (a) voiced speech with excitation generated by vocal cord vibrations, (b) unvoiced speech with excitation generated by turbulence in the vocal tract, and (c) mixed-mode speech, where both kinds of excitations are often present simultaneously.

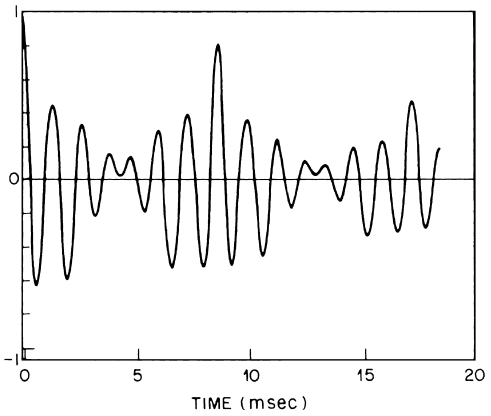


FIGURE 15 Autocorrelation function of voiced speech.

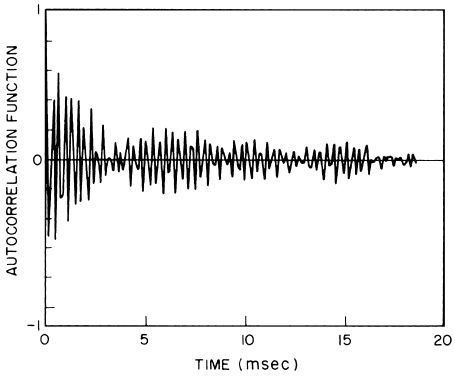


FIGURE 16 Autocorrelation function of unvoiced speech.

are not zero, although they are smaller compared to those of voiced speech shown in Fig. 15. The predictor parameters M and β for unvoiced speech are different from voiced speech; the value of β is generally much smaller and the value of M is likely to change randomly from one speech segment to another for unvoiced speech. Moreover, the random nature of unvoiced speech produces in general a higher prediction error than voiced speech.

You may wonder why we are using only one predictor coefficient in the pitch predictor. There is no good reason and the prediction can be extended to include several samples in the vicinity of the pitch period M. Let us now look at the prediction error after pitch prediction. An example is shown in Fig. 17, which shows the speech waveform and the prediction error after pitch prediction. As it should be, the prediction error does not show any obvious repetitive pattern.

Both predictors, one using speech samples in the immediate past (predictor with short delays) and another using speech samples in the previous pitch period (predictor with long delays), can be combined to take advantage of both kinds of prediction. The result is much improved prediction and very small prediction errors. The prediction error after two stages of prediction has very little resonance or pitch structure left and looks very much like a white noise signal. That is why the prediction error is often called the innovation signal, a signal representing only the new information. The speech signal and prediction error signals after both kinds of prediction are shown in Fig. 18.

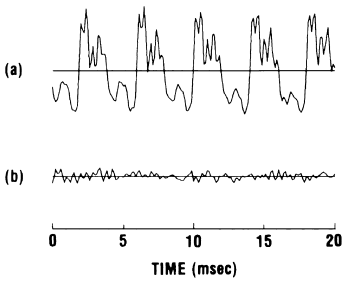


FIGURE 17 (a) Speech and (b) prediction error waveforms after pitch prediction.

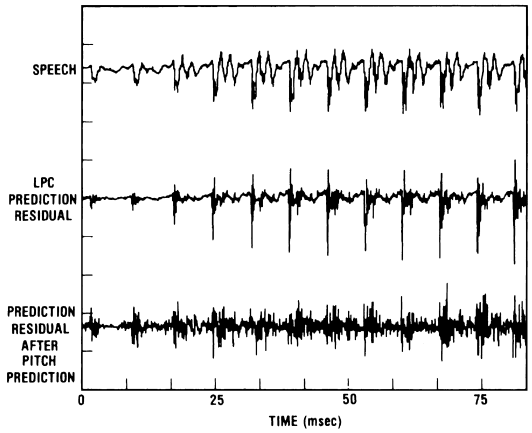


FIGURE 18 Speech and prediction residual (error) waveforms after prediction using short delays and using both short and long delays. Linear predictor coefficient (LPC) prediction residual (based on short delays) is amplified five times relative to the speech waveform. The prediction residual after pitch prediction is amplified by an additional factor of five.

Let us now briefly summarize the two predictions we have discussed. The first prediction uses only short delays in the prediction process, and we will denote this predicted value by $\hat{s}_1(n)$. The second prediction makes use of long delays in the prediction process and produces further reduction in the prediction error. The predicted value after both kinds of prediction based on short and long delays is denoted by $\hat{s}_2(n)$.

IV. APPLICATIONS OF LINEAR PREDICTION

We will now discuss some of the important applications of linear prediction analysis. The technique has been used for speech analysis and synthesis on computers, for digitizing speech for transmission on digital networks, and for speech and speaker recognition. We will discuss briefly some of these applications.

A. Speech Synthesis at Low Bit Rates

We are going to consider an important application of linear prediction—speech synthesis on computers. The synthesis procedure follows directly from Eq. (5), which can be rewritten as

$$s(n) = \varepsilon(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{21}$$

We find that the predictor coefficients can be quantized and represented accurately by a small number of bits, typically 50 bits for 10 coefficients. This implies an information rate

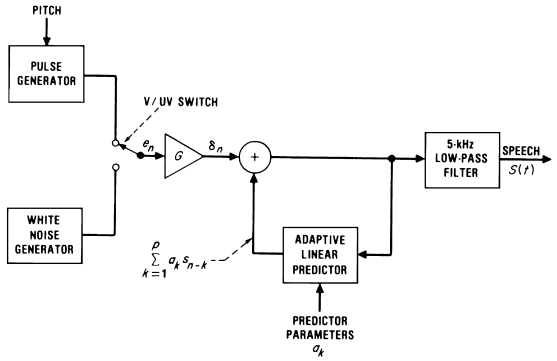


FIGURE 19 Low bit rate linear predictive speech synthesizer. The excitation is created by unit impulse train at pitch periods for voiced speech and white random noise for unvoiced speech. The speech energy is controlled by the gain factor G of the amplifier.

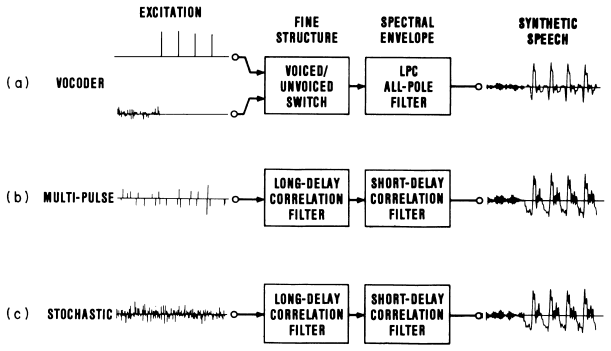


FIGURE 20 Three different speech synthesis models making use of linear predictive filters: (a) vocoder model using separate excitations for voiced speech and unvoiced speech, (b) multipulse model using a sequence of pulses, and (c) stochastic model using random noise. Both multipulse and stochastic models eliminate the need for voiced/unvoiced decision and pitch analysis.

of about 2500 bits/s for the predictor coefficients. To synthesize speech, we have to represent the prediction error $\varepsilon(n)$ at a low enough data rate, too. In the prediction error waveforms of Fig. 12, the prediction error looks like an impulse train. Thus for voiced speech, the prediction error $\varepsilon(n)$ could be replaced by an impulse train of adjustable amplitude and period. For unvoiced speech, it is usually sufficient to replace the prediction error by white random noise of adjustable amplitude. A speech synthesizer based on these ideas is shown in Fig. 19. The excitation control data consists of 1 bit for the voiced/unvoiced decision, and 6 bits for the pitch period. The amplitude of the speech output is controlled by the amplifier G, specified by 5 bits. The predictor coefficients require a total of 50 bits of 10 coefficients. Thus 62 bits per frame are required to characterize the speech. If these parameters are specified once every 20 ms, it results in a data rate of about 3000 bits/s—a considerable saving over the usual 64,000 bits/s for PCM.

The above synthesis procedure, often used in linear predictive vocoders, can produce speech of fairly good quality. The synthetic speech is usually intelligible but does not sound natural as spoken by a human talker. The reasons for the lack of naturalness are not exactly understood but they are primarily due to the replacement of the prediction error $\varepsilon(n)$ by a composite of pulses and white noise as shown in Fig. 19. Natural sounding speech of very high quality can be produced by using better representations of the prediction error. Three different synthesis models using linear predictive filters for speech synthesis are illustrated in Fig. 20.

The first synthesis model is the vocoder model discussed earlier. This model requires that speech be classified into two well-defined classes: voiced and unvoiced, and the pitch period for voiced speech be exactly known. As a practical matter, it is sometimes difficult to determine if speech is voiced or unvoiced. Also, since the speech signal is only approximately periodic it is very hard in some instances to locate the exact pitch period. These problems are completely avoided in the multipulse synthesis model shown in Fig. 20. The multipulse model makes no distinction between voiced or unvoiced speech nor does it require an exact knowledge of the pitch period. The prediction error is replaced by a sequence of pulses whose exact locations are determined by a matching procedure that minimizes the difference between the original and synthetic speech waveforms. Only a few pulses are needed to produce speech of very high quality. The small number of pulses (typically 4 pulses every 5 ms) translates into a small bit rate for representing the prediction error. However, the high speech quality of a multipulse synthesizer is achieved by a significant increase in the bit rate in comparison to a vocoder. A typical bit rate for the multipulse synthesizer is 8000 bits/s.

The stochastic synthesis model shown in Fig. 20 can reduce the bit rate for high-quality speech synthesis even further. In this model, the prediction error (input to the long-delay filter) looks very much like random stochastic noise; hence the name stochastic model. The best random approximation to the prediction error is selected from a prestored codebook of random sequences to minimize the difference between the original and synthetic speech signals, as we did in the multipulse model. Both multipulse and stochastic speech synthesis models can produce natural sounding speech of very high quality but the methods they employ for determining the best multipulse or random sequences are fairly complex and require a sophisticated computer to perform the necessary computations.

B. Speech and Speaker Recognition

Linear prediction analysis provides a set of predictor coefficients, typically 10, every 10 to 20 ms. This is a very

compact description of speech signals. Each set of coefficients could be considered a vector of predictor coefficients, and we will call them LPC vectors. Typically in speech recognition one compares a sequence of LPC vectors coming out of an unknown speech utterance with a number of stored reference templates that could represent either a sound, a word, a phrase, or even an entire sentence. Since a speech utterance can be spoken at different speaking rates, the comparison between the test utterance and the reference utterances must correct for these differences. Speech recognition using linear predictor coefficients works very well for isolated words and utterances with connected words.

Speaker recognition is similar to speech recognition, but here the task is to recognize or verify the identity of an unknown speaker. Again, one stores a set of reference templates, one for each speaker, and compares the test LPC vectors with the reference vectors after correcting for changes in the speaking rates between the two. Speaker recognition with LPC vectors is accurate, and error rates of less than 1% have been found.

SEE ALSO THE FOLLOWING ARTICLES

DIGITAL SPEECH PROCESSING • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, ANALOG • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL

BIBLIOGRAPHY

Atal, B. S., and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* **50,** 637–655.

Markel, J. D., and Gray, A. H., Jr. (1976). "Linear Prediction of Speech," Springer-Verlag, New York.

Rabiner, L. R., and Schafer, R. W. (1980). "Digital Processing of Speech Signals," Prentice Hall, Englewood Cliffs, N.J.



Spread-Spectrum Systems

Urbashi Mitra
Ohio State University

I. History
II. Multiple Access Methods
III. Spreading Sequences
IV. Signal Model
V. Basic Receiver Structures
VI. Performance Evaluation of Multiuser Receivers
VII. Wireless Channel Models
VIII. Additional Receiver Structures
IX. Channel Estimation
X. Power Control
XI. Conclusions

GLOSSARY

Asymptotic multiuser efficiency The relative decay rate of the probability of error of the receiver of interest in comparison to the single user optimal receiver operating in a single user environment.

Conventional detector The optimal detector for a single CDMA user in additive Gaussian noise; a matched filter detector matched to the signature waveform of the desired user. Suboptimal in the presence of multiple CDMA users.

Cross-correlation The inner product of two spreading waveforms. This parameter gives a measure of the distance between two spreading waveforms.

Effective energy Given a receiver with associated probability of error, the effective energy is the energy required in a single user system employing a matched filter receiver to achieve the same probability of error.

Multiple-access interference The noise due to interfering users during the detection of a particular desired user.

Multiuser detection terminology Formulae.

Multiuser efficiency The ratio of the effective energy to the transmitted energy of the user of interest.

Near–far problem The scenario where a desired user's signal is received at a power that is much less than that of the interfering users. Thus there is a possibility that an interferer's signal may overwhelm that of the desired user.

Near–far resistant receiver (heuristic) A receiver which is insensitive to differences in the received powers of interfering users.

Q-function Complementary cumulative distribution function for a standard Gaussian random variable.

Signature sequence The identifying code used by a user to (*spread*) its narrowband signal. Also known as a *spreading code* or (*spreading waveform*). These sequences are typically normalized to have unit energy.

THE RECENT EXPLOSIVE growth of wireless communications has motivated the "re-"consideration of spread-spectrum techniques for multi-user communications. As the phrase suggests, "multiuser" communications systems offer communication services to multiple users simultaneously. Our focus will be on a system as depicted by Fig. 1. In such a system, multiple users share a communications channel. The term "channel" is both abstract and physical; it describes the link between the transmitter(s) and the receiver(s). Thus, it could refer to freespace or even a body of water. For free space, the channel is typically defined by a band of frequencies and characterized by the physical topology between the transmitter(s) and the receiver(s). This multiuser system can also be termed a *multipoint-to-point* communications system in contrast with a *point-to-multipoint* or *broadcast* system as employed for broadcast radio transmission and broadcast television. In broadcast channels, a single information stream is transmitted from a centralized transmitter to be received by multiple receivers. The objective of the receiver in our multipoint-to-point or *multiple access* system is to ultimately demodulate the information stream of one, some or all of the active users in the system. The receiver is thus the recipient of the different information signals of multiple users. Examples of multiuser communication systems include cellular communications, local area networks, computer communications networks (such as the Internet), telephone networks, and packet-radio networks. In Fig. 1 the additive noise due to the channel is distinguished from the contributions of the individual users. However, each active user in the system can represent a noise source for every other user in the system. The challenge of designing a receiver which operates well in a multiuser environment is to mitigate the effects of both the interfering users as well as the effects inherent to the wireless channel due to propagation and ambient channel noise.

Spread-spectrum signaling while somewhat bandwidth inefficient relative to more traditional narrowband signaling schemes offers certain advantages for radio communication systems. The wideband nature of the signal facilitates channel estimation and enables the resolution of multipath (described in more detail in Section VII.A). Multipath occurs to the presence of obstructions such as buildings, trees, and vehicles in the path between transmitter and receiver. Due to these obstructions, the transmitted signal is reflected, absorbed, and diffused; the received signal is in fact a sum of delayed and attenuated replicas of the originally transmitted signal. The access schemes associated with spread spectrum technology tend to be more flexible; and statistical multiplexing can be exploited since many of the active users have bursty communication. Thus, there is potential for a capacity increase relative to narrowband signaling systems.

We shall focus on adaptive schemes for data detection; however, adaptive algorithms can also be developed for adaptive estimation of key communication parameters such as the channel, the number of active users, timing information, etc.

I. HISTORY

Prior to describing the details of spread-spectrum multiuser systems, we note some of the relevant events of the past. The history of spread-spectrum communications has close connections to the birth of information theory, a Hollywood starlet, and the necessities and mores of

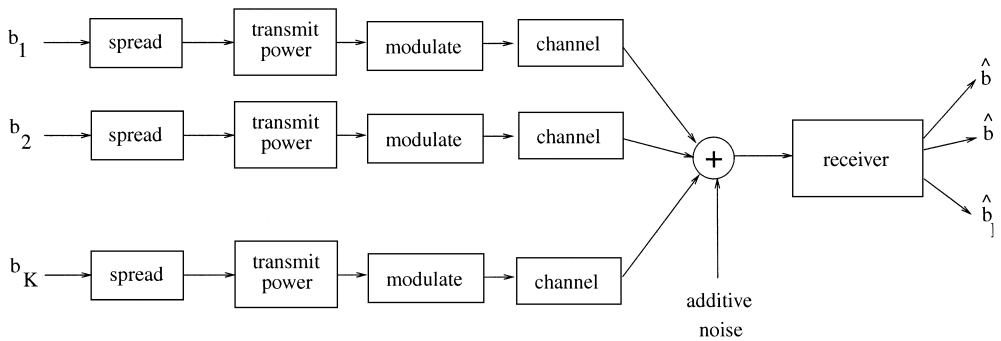


FIGURE 1 Multiuser system.

war-motivated research.[1] Essential to the idea of communicating via a scheme that employs more resources than necessary is the notion of *bandwidth* or transmission rate. As early as the 1920s, Nyquist (1924) began to quantify this concept in his work on telegraph speed. The goal of quantifying information rate continued in 1928 with work by Hartley, who endeavored to "set up a quantitative measure whereby the capacities of various system to transmit information may be compared." This continued effort culminated in the publication of the seminal work by Claude Shannon, "A Mathematical Theory of Communication," which heralded the birth of information theory and showed the existence of methods by which error-free communication could occur provided that the communication did not exceed an information rate defined by the communication channel. Shannon's subsequent work on the capacity of a channel whose only distortion is additive white Gaussian noise further motivated the consideration of wideband signaling to combat jamming.

Prior to the birth of information theory, technologists were considering the benefits of spreading the bandwidth of an information signal to improve fidelity at the receiver. Armstrong's work on frequency modulation (FM) in the 1930s for radio revealed that the combination of spectrum spreading with preemphasis and deemphasis filtering at the transmitter and receiver, respectively, could improve immunity to noise over traditional single-side band modulation (see Proakis and Salehi, 1994). Shortly thereafter, radar engineers were recognizing the benefits of wideband signaling to improve time delay estimation and hence ranging. In 1938, Guanella filed for a radar patent where the transmitted signal was "composed of a multiplicity of different frequencies, the energies of which are small compared to the total energy."

World War II spawned much communications research on spread-spectrum type communications. Given the proliferation of radio and radar systems to aid in communications and weapons guidance, there was a need to develop systems which would be immune to enemy jamming. With narrowband communications, it was possible to employ a radiometer to determine the carrier frequency being employed and then simply transmit a strong jamming signal at the same carrier to drown out the signal. Perhaps one of the most popularized efforts toward anti-jam communications during this time was the patent filed by Hedy Lamarr, the Hollywood actress, and her composer/inventor colleague, George Antheil in 1941. The patent proposes the use of a sequence of frequencies for communication. The pattern of frequencies is to be known at the transmitter and receiver and stored on player-piano-like rolls. Without the frequency sequence, it is near impossible to demodulate or jam the transmission. This patent presages spread-spectrum systems based on frequency hopping.

The postwar era resulted in much work on spread spectrum communications including the production and storage of noise-like waveforms, direct-sequence and frequency hopped systems, the introduction of the notion of processing gain. The bulk of this research, however, was classified and conducted under the auspices of the Department of Defense (Scholtz, 1994).

While it may not be possible to determine the origins of the multiple-access by users asynchronously accessing the channel, as early as 1949, a technical memorandum by Pierce (1949) discussed this notion with credit to Shannon. An integral part of this discussion was the realization that more and more users could enter the communication channel with "graceful degradation" in performance so long as the waveforms employed by the users were near-orthogonal. It was observed by White, in the same year, in his work on satellite systems that these asynchronous multiplexing techniques were most appropriate for spread-spectrum systems.

In a celebrated exchanged between Price and Shannon, the idea of code-division multiple-access is likened to a party,

> . . . More and more people can come, and they would all pay equally, so to speak. If more people were there, gradually the noise level would increase on each channel. But everyone could still talk, even though it might be a pretty noisy 'cocktail party' by that time.

This observation of graceful performance loss is informed by the technology of the time; that is, the receiver employed in code-division multiple access schemes was a simple correlator matched to the spreading waveform of the user of interest. As will be seen in the sequel, such a receiver is suboptimal and suffers when there are users in the system that are received more powerfully than the user of interest. In fact, this effect, dubbed the *near–far effect* was noted as early as 1961 in a paper by Magnuski (1961). Processing gain, or the amount of bandwidth expansion induced by spreading is discussed in De Rosa and Rogoff (1950).

In the 1960s and 1970s, spread-spectrum technology was advanced through work on military satellite communications. Power control and power distribution problems were addressed (Scholtz, 1994). The ranging capabilities of spread-spectrum systems are exploited in Global Positioning Systems; these systems were introduced in the mid 1970s. While initially developed as a military navigation aid, such systems are currently commercially available.

[1]This historical discussion is based, in bulk, on the excellent article by Scholtz (1994). A more extensive history is available in Scholtz (1982).

The commercial use of spread-spectrum techniques is closely tied to the birth of modern cellular communication systems. The first mobile radio systems were for public safety (Chapin, 1960), in the early 1900s. The Advanced Mobile Phone Service (AMPS) deployed in 1978 was based on a frequency-division multiple-access (FDMA)[2] method was the first public radio telephony communication system in the United States. However, even at that time, it was hypothesized that spread-spectrum techniques might offer improved performance (Cooper and Nettleton, 1978). AMPS and related services of the time are called *first-generation* cellular services. Interestingly, Hewlett-Packard had constructed an indoor wireless product based on spread-spectrum methods (Freret, 1980), but did not pursue its commercialization. This historical fact is particularly interesting given that the IEEE 802.11 standard for indoor wireless local area networks exploits spectrum spreading to reduce power transmission. Today, there are numerous commercial products operating to the 802.11 standard.

The first-generation systems such as AMPS, were analog communications systems. As the technology advanced and digital communication systems proliferated, digital systems were considered for wireless communications as well. As a result, both time-division multiple-access (TDMA) and code-division multiple-access (CDMA) methods were pursued. The study of spread-spectrum systems was facilitated by the Federal Communications Commission's 1985 policy which encouraged experimental work on spread-spectrum. The first spread-spectrum CDMA set of interim standards was put forth in Interim Standard 95 (IS-95) in 1993. In January, 1994, the company Qualcomm field-tested some of the first CDMA base stations in Seattle, Washington. Prior to these tests, Qualcomm had another CDMA product—Omnitracs, a truck-tracking/communications system first sold in 1988. Since the introduction of IS-95 in 1993, several revisions have occurred. The CDMA Personal Communications System (PCS) is in the 1.8- and 2.0-GHz bands. These current systems are considered to be second generation wireless systems.

In the current work, we shall focus on Direct-Sequence CDMA (DS-CDMA). This focus is motivated by the imminent adoption of the DS-CDMA type signaling in a variety of third-generation wireless standards (Adachi *et al.*, 1998; Dahlman *et al.*, 1998; Ojanperä and Prasad, 1998). It is observed that for both the frequency division duplex (FDD) and the time division duplex (TDD) modes of UMTS, DS-CDMA multiple access is laid over the FDD and TDD duplexing schemes (Haardt *et al.*, 2000).

[2]Multiple access methods are detailed in Section II.

As a concluding note to the discussion of multiple access schemes, we observe that TDMA and FDMA are special cases of CDMA, where typically the "spreading waveforms" are mutually orthogonal. Thus, with proper signal description, the methods described herein have utility for TDMA and FDMA systems where there is co-channel or adjacent interference caused by dispersion or insufficient frequency reuse.

II. MULTIPLE ACCESS METHODS

As noted previously, we are interested in multipoint to point systems as typified by a base-station communicating with multiple mobile stations. The emphasis of this chapter is on signaling and detection methods appropriate for multiuser radio communications; however, it is observed that multiuser demodulation methods have found application in a variety of other fields including radar signal processing and medical imaging. There are multiple ways in which the radio channel resource can be shared. Two more classical methods are Frequency-Division Multiple-Access (FDMA) and Time-Division Multiple-Access (FDMA) (see Fig. 2). For illustrative purposes, one can view the communications resource as having two dimensions: frequency and time. In reality, other dimensions are available such as space.

The first wireless mobile communications system, AMPS, employed FDMA as the multiuser access technology in 1977. In FDMA, each user is assigned a frequency band; the user can communicate continuously employing this frequency "slot." Commercial radio and broadcast television employ FDMA as a method of transmitting many different station signals to an individual receiver. In TDMA, each active user is assigned a nonoverlapping time slot. During its assigned time slot, the active user transmits over the entire frequency band allocated to the TDMA service. The TDMA access scheme can be considered to be the dual of FDMA: users are assigned nonoverlapping time slots and utilize the entire frequency band during this time slot. Users share the resource by communicating one-at-a-time, in round-robin fashion. Currently, there are three variations of TDMA in use for commercial wireless communications. The Global System for Mobile Communications (GSM) is widely deployed in Europe, North America, and parts of Asia in the 900-, 1800-, and 1900-MHz frequency bands. In place in Japan is the Pacific Digital Cellular system which also employs TDMA. And finally, North America Digital Cellular exists in North America at the 800- and 1900-MHz bands.

Both of these methods, FDMA and TDMA, assign orthogonal channels to each active user and both methods have a maximum number of users that can be serviced

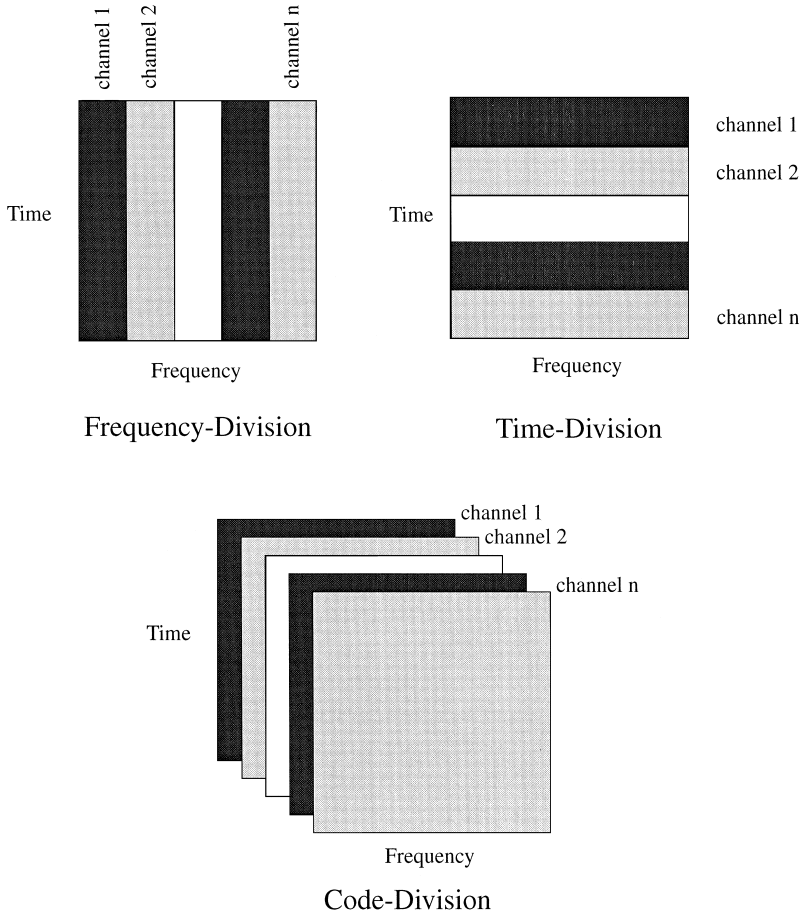


FIGURE 2 Multiple-access methods.

simultaneously. We thus consider TDMA and FDMA to be fixed resource allocation schemes. We note that if there are fewer users than the maximum number of "slots," resources can be wasted (not used) for these fixed resource allocation schemes.

The multiple-access strategy of interest herein is Code-Division Multiple-Access (CDMA), also illustrated in Fig. 1. In CDMA, each active user is assigned a waveform which exploits the total time and frequency bands allocated for the service. Both TDMA and FDMA can be considered as special cases of CDMA. By allowing for user waveforms with more general characteristics, CDMA signals can be designed to be more immune to the effects of the wireless channel and to allow for more users to share the radio channel. This additional user capacity, however, will come at the expense of degradation in performance or at the expense of a more sophisticated and thus generally more complex receiver structure.

The particular type of CDMA that shall be considered is short code Direct-Sequence CDMA (DS-CDMA). In DS-CDMA, the waveform assigned to each user is generally designed to have a relatively flat spectrum and thus can be viewed as almost random. We next describe some of the basic features of spreading waveforms and how they are generated.

III. SPREADING SEQUENCES

In DS-CDMA, the waveform assigned to each user can be viewed as being a function of a *spreading sequence* and a pulse waveform. The spreading sequence, $\mathbf{s}$, is drawn from a finite alphabet (e.g., $\{\frac{\pm 1}{\sqrt{N}}\}$) and modulated onto a pulse shape $p(t)$ (e.g., a rectangular pulse shape or a raised cosine pulse shape). To provide a consistent definition of signal-to-noise ratio per bit, the spreading waveforms are normalized $\|\mathbf{s}_k\|^2 = 1$. The parameter N is the length of the spreading sequence and is also known as the *processing gain.* It is a measure of the bandwidth expansion offered by the spreading operation. The distinction of "short code" DS-CDMA indicates that we shall be focusing on systems where the same spreading sequence is used for each bit transmitted. This is in contrast to the past and present implementations of standardized DS-CDMA, where the

spreading sequence is time-varying and has a period that is equal to many symbol intervals. An example,

$$\mathbf{s} = \frac{1}{\sqrt{N}}[-1,-1,+1,-1,+1,-1,+1,+1]$$

$$p(t) = \begin{cases} 1 & t \in [0, T_c) \\ 0 & \text{else} \end{cases}.$$

The parameter T_c is called the chip duration and the symbol duration is thus $T = NT_c$.

The spreading sequences are chosen to have desirable auto-correlation and cross-correlation properties (Sarwate and Pursley, 1980). Several key sets of spreading sequences are derived from maximal length shift register sequences, also known as m-sequences. The m sequence is generated by a linear feedback shift register whose operation is governed by a primitive polynomial of degree m,

$$p(x) = 1 + p_1 x + p_2 x^2 \cdots p_{m-1} x^{m-1} + p_m x^m,$$

the coefficients are drawn from $\{0, 1\}$. Such a system, with properly chosen coefficients, will yield a sequence of length $N = 2^m - 1$. The desired polynomials are tabulated in a variety of texts (see, e.g., Sarwate and Pursley, 1980; Simon *et al.*, 1985; Ziemer and Peterson, 1985). The resulting $\{0, 1\}$ sequence is mapped to a $\{1, -1\}$ sequence and normalized. The cyclic autocorrelation function of the sequence is defined to be

$$\rho(\tau) = \sum_{i=1}^{N} \underline{s}(i)\underline{s}((i+\tau) \bmod N),$$

where $\underline{s}(i)$ denotes the ith component of the vector sequence $\underline{s}$ and mod is the modulo operator. The generated m-sequences have the following properties:

1. The number of negative elements exceeds the number of positive elements by one.
2. The autocorrelation function is given by

$$\rho(\tau) = \begin{cases} 1 & \tau = lN, l = 0, 1, \ldots \\ -\frac{1}{N} & \text{else} \end{cases}.$$

A key disadvantage of m-sequences in regards to commercial spread-spectrum systems is that the number of distinct m-sequences of a certain length that can be generated is quite limited.[3] Furthermore, pairs of m-sequences do not always achieve the most desirable cross-correlation functions. Two sets of pseudo-noise sequences which do have good cross- and auto-correlation functions are Gold (1967) sequences and Kasami sequences (Kasami *et al.*, 1968).

Gold sequences are constructed by taking the sum of $\underline{s}_1$ with all cyclic shifts of $\underline{s}_2$, where $\underline{s}_1$ and $\underline{s}_2$ are two m-sequences selected such that their associated primitive polynomials have a key relationship (Gold, 1967; Stüber, 1996). The resulting set of sequences is of cardinality $2^m + 1$ and furthermore, it can be shown that the resultant auto- and cross-correlation functions are three valued, where those three values are quite modest. Kasami sequences are constructed in a similar fashion to Gold sequences; however, the constituent sequences have period $2^m - 1$ and $2^{\frac{m}{2}} - 1$, respectively; the resultant spreading sequence cardinality is $2^{\frac{m}{2}}$. Kasami sequences also have three-valued auto- and cross-correlation functions; however, these values are smaller in magnitude than those of Gold sequences.

Another set of sequences, which are not pseudo-noise sequences, and possess poor off-peak correlation values are the Walsh–Hadamard sequence sets. This set is worthy of mention due to its prominent role in second generation CDMA standards (Rappaport, 1996; Viterbi, 1995). These sequences occur in the spreading operations and in the modulation specified for these standards. The rows of the Hadamard matrix of dimension N constitute the desired sequences. The Hadamard matrices are determined using the following recursion:

$$\mathbf{H}_{2N} = \begin{bmatrix} \mathbf{H}_N & \mathbf{H}_N \\ \mathbf{H}_N & -\mathbf{H}_N \end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix} +1 & +1 \\ +1 & -1 \end{bmatrix}.$$

It is straightforward to show that the rows of the Hadamard matrix are orthogonal. Thus, if used in a truly synchronized system, in the absence of frequency selective fading (propagation effects), Hadamard spreading sequences result in optimal receivers of low complexity (see Section V, which details several key receiver structures for CDMA systems). However, this is an idealized scenario that rarely occurs in practice.

IV. SIGNAL MODEL

For illustrative purposes, we will initially focus on a simplified communication scenario: bit synchronized multiple users communicating over an additive white Gaussian noise channel. Generalizations of this model are considered in Section VII.A. We shall assume that the front-end filter of the receiver is synchronized to the users and furthermore is coherent. The chip-matched filtered signal can be represented as a sequence of vectors:

[3] This number is given by $\Phi(2^m - 1)/m$ for a length $2^m - 1$ sequence where $\Phi(n) = n\Pi^n p \,|\, n(1 - \frac{1}{p})$ is the Euler totient function and $p \,|\, n$ denotes the prime numbers p which divide n.

$$\mathbf{r}(i) = \sum_{k=1}^{K} A_k b_k(i)\mathbf{s}_k + \mathbf{n}(i)$$
$$= \mathbf{SAb}(i) + \mathbf{n}(i),$$

where

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K]$$
$$\mathbf{A} = \text{diag}[A_1, A_2, \ldots, A_K]$$
$$\mathbf{b}(i) = [b_1(i), b_2(i), \ldots, b_K(i)]^T,$$

where K is the number of active users and $\mathbf{n}(i)$ is the additive noise process, modeled as a white, Gaussian random vector process with zero-mean and covariance $\sigma^2\mathbf{I}_N$. The matrix $\mathbf{I}_L$ is an $L \times L$ identity matrix. User k's received amplitude, data bit at time i and spreading sequence are denoted by A_k, $b_k(i)$, and $\mathbf{s}_k$, respectively. Herein, we shall assume binary-phase shift-keyed (BPSK) data; that is, $b_k(i) = \pm 1$. If required, the statistical model for the data is that $b_k(i)$ takes on its binary values with equal probability ($\frac{1}{2}$).

To facilitate the description of some classical multiuser receivers, we introduce the following notions. It can be shown that a set of sufficient statistics for detecting the data of a single user or all users in a multiuser CDMA system is the output of a bank of filters matched to the spreading code of each active users. This fact about multiuser signaling is key to the development of multiuser receivers. A statistic $T(\mathbf{r})$ is sufficient for estimation of a parameter θ if the probability density function of the observation, $\mathbf{r}$ conditioned on the statistic, $p(\mathbf{r} \mid T(\mathbf{r}))$ is not a function of the desired parameter θ (Poor, 1994; Porat, 1994). All matched filter outputs and not just the matched filter output associated with the user of interest are sufficient for data estimation in a multiuser system. This fact alludes to the suboptimal performance of the matched filter receiver.

These matched filter outputs are defined as,

$$\mathbf{y}(i) = \mathbf{S}^H\mathbf{r}(i) = \mathbf{RAb}(i) + \tilde{\mathbf{n}}(i),$$

where

$$\mathbf{R} = \mathbf{S}^H\mathbf{S}.$$

The noise present in the system is now colored, $\tilde{\mathbf{n}}(i) \sim \mathcal{N}(\mathbf{0}_K, \sigma^2\mathbf{R})$, where $\mathbf{0}_L$ is a $L \times 1$ vector of zeros. The $K \times K$ cross-correlation matrix $\mathbf{R}$ can be interpreted as a catalog of how far apart pairs of spreading codes are in the multiuser system. The Euclidean distance between two spreading codes is

$$\| \mathbf{s}_j - \mathbf{s}_k \|^2 = \| \mathbf{s}_j \|^2 + \| \mathbf{s}_k \|^2 - 2\mathbf{s}_j^H\mathbf{s}_k$$
$$= 2 - 2\mathbf{R}[j, k].$$

We shall see that the performance of any multiuser or single user receiver is very dependent on the values of the components of $\mathbf{R}$. Note that a set of K mutually orthogonal spreading codes leads to $\mathbf{R} = \mathbf{I}_K$.

Another key matrix for multiuser detection is the data correlation matrix, that is

$$\mathbf{C} = \mathbf{E}\{\mathbf{r}(i)(i)r^H\} = \mathbf{SA}^2\mathbf{S} + \sigma^2\mathbf{I}_N. \quad (1)$$

The operator $\mathbf{E}\{\cdot\}$ denotes expectation over all random quantities in the argument, unless otherwise specified.

V. BASIC RECEIVER STRUCTURES

We next briefly review several important multiuser receivers: the conventional receiver, the decorrelating detector, the minimum-mean squared error receiver, the jointly optimal detector, the individually optimal detector, and interference canceling receivers. These receivers vary in their offered performance and attendant complexity. Multiuser detection schemes which exploit the fact that the received interference is both structured (multiple-access interference) as well as unstructured (additive Gaussian noise) offer superior performance to those receivers which do not. This observation is motivated by the sufficiency arguments in the previous section. An alternative perspective is the modernization of Shannon's cocktail party analogy: in a multiuser DS-CDMA system, we have an *international cocktail party*.[4] As more and more party attendees arrive, the ambient noise level rises; however, if we further utilize the fact that not all participants speak the same language, we can actually increase the size of the party and still have communication with reasonable fidelity. Good multiuser detection schemes exploit the different languages (different spreading codes) in the system, either explicitly or implicitly.

We additionally note that the history of multiuser detection design is a further history lesson on methods for equalization of inter-symbol interference channels in single-user communication systems (add in citations for key stuff).

A. Conventional Receiver

This is the classical receiver which was considered optimal prior to a deeper understanding of multiple access interference (MAI). It was originally argued that the Central Limit Theorem (see, e.g., Papoulis, 1984; Stark and Woods, 1994) held and thus the MAI could be modeled as additive, white, Gaussian noise (see, e.g., Pickholtz *et al.*, 1982; Simon *et al.*, 1985). The optimal single user receiver

[4]This analogy is attributed to Geoffrey Orsak.

for transmission in additive white Gaussian noise is the matched filter receiver. The conventional or matched filter receiver output is given by

$$\hat{\mathbf{b}}(i) = \text{sgn}(\mathbf{y}(i)), \tag{2}$$

where the operator $\text{sgn}(\cdot)$ outputs the sign of each component of its argument. The conventional receiver is MAI limited and incurs a high probability of error if the power of the received signal of the desired user is significantly less than that of the interfering users. This undesirable property of the conventional receiver is termed the **near far problem**. We note that if a receiver is insensitive to the near far problem it is deemed *near far resistant.* The conventional receiver also suffers if the system is highly loaded. We note that there exists one scenario in which the conventional receiver is actually optimal in terms of probability of bit error—this occurs when the spreading codes of the active users are mutually orthogonal. In a realistic wireless system, this property is difficult to maintain. It is noted, however, that the conventional receiver is very straightforward to implement and only requires knowledge of the desired user's spreading waveform and timing.

B. Decorrelating Detector

The decorrelating detector (Lupas and Verdú, 1989, 1990; Schneider, 1979) is the receiver which zero-forces the MAI. That is to say, it completely nulls out the MAI at the possible expense of removing some of the signal energy of the user of interest. It can also be viewed as the maximum-likelihood estimator of the vector $\mathbf{Ab}(i)$. This receiver is formed by

$$\hat{\mathbf{b}}(i) = \text{sgn}(\mathbf{R}^{-1}\mathbf{y}(i)). \tag{3}$$

The direct construction of the decorrelator requires the knowledge of the spreading waveforms of all of the active users and the associated timing information. Despite the simplicity of this receiver, the decorrelator shares many properties with that of the optimal detector to be discussed in the sequel.

C. Minimum Mean-Squared Error Receiver

To introduce the the linear minimum mean-squared error (MMSE) receiver (Madhow and Honig, 1994; Xie *et al.*, 1990), we first discuss a generic linear receiver. Let $\mathbf{z}(i)$ be the soft-output of a general linear receiver $\mathbf{M}$,

$$\mathbf{z}(i) = \mathbf{M}\mathbf{r}(i) \tag{4}$$

$$\hat{\mathbf{b}}(i) = \text{sgn}(\mathbf{z}(i)). \tag{5}$$

Then, the MMSE receiver is determined by

$$\mathbf{M} = \arg\min_{\mathbf{M}} \mathbf{E}\{\|\mathbf{b}(i) - \mathbf{z}(i)\|^2\}. \tag{6}$$

Two equivalent solutions, ignoring positive scalings, are given by

$$\mathbf{M} = \mathbf{S}^T\left(\mathbf{S}\mathbf{A}^2\mathbf{S}^T + \sigma^2\mathbf{I}_N\right)^{-1} \tag{7}$$

$$= (\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}\mathbf{S}^T. \tag{8}$$

The two forms of the MMSE receiver can be shown to be equivalent through the use of the matrix inversion lemma (Golub and Van Loan, 1989; Horn and Johnson, 1985). Thus, the MMSE estimate of the data is given by

$$\hat{\mathbf{b}}(i) = \text{sgn}((\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}\mathbf{y}(i)). \tag{9}$$

In addition to the information required by the decorrelating detector, the MMSE receiver is also a function of the received amplitudes of the active users and the noise variance. In general, the MMSE receiver outperforms the decorrelating detector. As will be observed in the following, decentralized implementations of the MMSE receiver (and decorrelator) exist. In considering these decentralized receivers, a few observations can be made in regards to asymptotic behavior. As the noise variance grows ($\sigma^2 \to \infty$) or as the interfering amplitudes diminish ($A_2, \ldots, A_K \to 0$), the MMSE receiver approaches the conventional receiver. Alternatively, as the noise variance decreases ($\sigma^2 \to 0$), the MMSE receiver converges to the decorrelating detector. The MMSE receiver performs the optimal tradeoff in combating multiple access interference versus suppressing ambient channel noise in a linear receiver.

D. A Few Points on Linear Receivers

We note that the prior receiver algorithms are all linear in nature. To summarize, the bit decision for a particular user, say user 1, can be written as

$$\hat{b}_1(i) = \text{sgn}\left(\mathbf{c}_1^H\mathbf{r}(i)\right). \tag{10}$$

Note that for data demodulation, a positive scaling of the receiver does not affect the decision. Thus for $\alpha > 0$, both $\mathbf{c}_1$ and $\alpha\mathbf{c}_1$ yield the same decision. For the sequel, we shall note the soft decision statistic for user k as

$$z_k(i) = \mathbf{c}_k^H\mathbf{r}(i). \tag{11}$$

Thus for the joint receivers discussed above we can define the following single user (decentralized) linear receiver vectors,

1. Conventional receiver $\mathbf{c}_k = \mathbf{s}_k$.
2. Decorrelating detector $\mathbf{c}_k = \mathbf{S}\underline{\rho^{-1}}_k$, where $\underline{\rho^{-1}}_k$ the kth column of $\mathbf{R}^{-1}$.

3. MMSE receiver $\mathbf{c}_k = \mathbf{S}\mathbf{m}_k$, where $\mathbf{m}_k$ the kth column of $(\mathbf{R} + \sigma^2\mathbf{A}^{-2})^{-1}$.

We next consider nonlinear receivers. These receivers are sometimes more complex to implement than the previously discussed linear receivers, the benefit of this added complexity is improved performance.

E. Jointly Optimal Detector

The jointly optimal multiuser receiver is formed by determining the maximum likelihood estimate of $\mathbf{b}$ (Verdú, 1986, 1998). Thus,

$$\hat{\mathbf{b}}(i) = \arg\max_{\mathbf{b}} p(\mathbf{y}(i) \mid \mathbf{b}) \tag{12}$$

$$= \arg\max_{\mathbf{b}(i)} 2\mathbf{b}(i)^H\mathbf{A}\mathbf{y}(i) - \mathbf{b}(i)^H\mathbf{A}\mathbf{R}\mathbf{A}\mathbf{b}(i) \tag{13}$$

$$= \arg\max_{\mathbf{b}(i)} \Omega(\mathbf{b}(i)), \tag{14}$$

where

$$\Omega(b) = 2\mathbf{b}^T\mathbf{A}\mathbf{y} - \mathbf{b}^T\mathbf{A}\mathbf{R}\mathbf{A}\mathbf{b}. \tag{15}$$

The spreading waveforms and amplitudes of all active users are necessary to construct the jointly optimal receiver. Note that the decorrelating detector only requires the spreading waveforms of the active users and not the amplitudes.

F. Individually Optimal Detector

The individually optimum receiver achieves the minimum probability of error for the user of interest (Verdú, 1986, 1998). Without loss of generality, we shall assume that user 1 is the desired user. Then, the individually optimum receiver is obtained by

$$\hat{b}_1(i) = \operatorname{sgn}\left[\sum_{\mathbf{b},b_1=1} \exp\left(\frac{\Omega(\mathbf{b})}{2\sigma^2}\right) - \sum_{\mathbf{b},b_1=-1} \exp\left(\frac{\Omega(\mathbf{b})}{2\sigma^2}\right)\right]. \tag{16}$$

The individually optimal detector requires the same set of parameter knowledge as the MMSE detector: user spreading waveforms, amplitudes, the noise variance, and timing. We observe that the individually optimal detector converges to the jointly optimum receiver as $\sigma \to 0$.

G. Interference Cancellation Receivers

There is a class of receivers which employs several iterations or stages to refine data bit estimates. A subset of these receivers can be called *interference cancellation* receivers as they endeavor to estimate the contribution to the received from all or some of the active users. The simplest interference cancellation receiver is the conventional receiver based successive interference cancellation receiver:

$$\hat{b}_k = \operatorname{sgn}\left(y_k - \sum_{j=1}^{k-1} A_k\hat{b}_j\rho_{jk}\right).$$

The decision of the first bit is simply the conventional receiver output. This interference cancellation receiver works best when the users are ordered from strongest (user 1) to weakest (user K). This is due to the fact it is necessary for the first users to be detected with high fidelity in order to ensure good performance for subsequent users (Patel and Holtzman, 1994). A more sophisticated successive interference cancellation scheme can be constructed by exploiting the Cholesky decomposition of the spreading code correlation matrix $\mathbf{R} = \mathbf{F}^T\mathbf{F}$ (Horn and Johnson, 1985). The matrix $\mathbf{F}$ is lower triangular. It can be shown that the linear transformation $\mathbf{F}^{-T}$ is a noise whitening transformation for the matched filter output vector. That is,

$$\mathbf{F}^{-T}\mathbf{y} = \bar{\mathbf{y}} = \mathbf{F}\mathbf{A}\mathbf{b} + \bar{\mathbf{n}}$$

where

$$\bar{\mathbf{n}} \sim \mathcal{N}\left(\mathbf{0}_K, \sigma^2\mathbf{I}_K\right).$$

We note that $\bar{\mathbf{y}}_1 = \mathbf{F}_{11}A_1b_1 + \bar{\mathbf{n}}_1$ and furthermore that $\mathbf{F}_{11} = 1/\sqrt{\mathbf{R}_{11}^{-1}}$. Thus the estimate of the first bit from $\bar{\mathbf{y}}_1$ is equivalent to the decorrelator estimate of the first bit. The successive interference cancellation receiver is given by

$$\hat{b}_k = \operatorname{sgn}\left(\bar{y}_k - \sum_{j=1}^{k-1} A_k\hat{b}_j\mathbf{F}_{jk}\right).$$

This receiver is deemed the *decorrelating decision feedback receiver* (Duel-Hallen, 1992) and can achieve near single-user performance for the final user in fact $\mathbf{F}_{KK} = 1$. This receiver does not require ordering of the users (note that the first bit's soft statistic is devoid of MAI); however, performance is improved if ordering is performed.

The previous two receivers detected bits in a serial fashion; it is also possible to do multi-stage detection in a parallel fashion. A parallel interference cancellation scheme is specified,

$$\hat{\mathbf{b}}^{m+1}(i) = \operatorname{sgn}\left[\mathbf{y} - (\mathbf{R} - \mathbf{I}_K)\mathbf{A}\hat{\mathbf{b}}^m(i)\right].$$

The superscript m indicates the stage of the estimates. There is some choice as to how to initialize the estimation scheme, one can use conventional decisions or decorrelated decisions (Varanasi and Aazhang, 1990).

The appeal of these nonlinear receivers is their often near-single user performance with attendant moderate implementation complexity. It should be noted, however,

that they can be quite sensitive to mismatched amplitude estimation information and do not operate very well in low SNR regimes.

VI. PERFORMANCE EVALUATION OF MULTIUSER RECEIVERS

For digital communications, the desired performance evaluation metric is typically the probability of bit detection error. Due to the nature of multiuser spread-spectrum signals and their associated detectors, computing this measure is often problematic. To this end, some more tractable measures have been developed which capture a receiver's ability to combat MAI. It should be emphasized that the probability of error of a particular receiver is dependent on many quantities:

- The noise variance
- The spreading codes of all active users
- The amplitudes of all active users
- The receiver itself

To begin, we note that for a single-user channel,

$$\mathbf{r} = A\mathbf{s}b + \mathbf{n}, \tag{17}$$

the optimal receiver is the matched filter which takes the sign of the following statistic $y = \mathbf{s}^H\mathbf{r}$. The statistic y has a Gaussian distribution, $y \sim \mathcal{N}(A, \sigma^2)$ where the original noise vector $\mathbf{n} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2\mathbf{I}_N)$. The resulting probability of bit detection error is thus

$$P(\sigma) = Q\left(\frac{A}{\sigma}\right), \tag{18}$$

where $Q(x)$ is the complementary cumulative distribution function[5] of a zero-mean, unit-variance Gaussian random variable. With the performance of a single user system completely characterized, we can discuss multiuser systems. We assume that the phenomena we wish to quantify is a receiver's ability to suppress MAI. To this end, the effects of the additive Gaussian noise are of secondary importance. We will observe that there are situations where increasing the transmitted power of a particular user may not be sufficient to improve the performance at the receiver.

Qualitatively, the *asymptotic multiuser efficiency* (Verdú, 1986) measures the asymptotic amount of power necessary for a multiuser receiver to emulate the performance of a corresponding single-user receiver as the ambient Gaussian noise diminishes. If a particular multiuser detector offers the probability of error $P_k(\sigma)$ for the detection of user k in the presence of additive Gaussian noise with variance σ^2, we can define the *effective energy*, $e_k(\sigma)$ for user k as,

$$e_k(\sigma) = \left\{\gamma \geq 0\colon P_k(\sigma) = Q\left(\frac{\sqrt{\gamma}}{\sigma}\right)\right\},$$

recall that $Q(\sqrt{\gamma}/\sigma)$ is the probability of detection error of a single user with received amplitude $\sqrt{\gamma}$ communicating only in the presence of ambient Gaussian channel noise with variance σ^2. Thus $e_k(\sigma)$ is a measure of the performance loss due to the presence of the interfering users in a particular multiuser Gaussian channel. The asymptotic efficiency is simply defined as

$$\eta_k = \lim_{\sigma\to 0}\frac{e_k(\sigma)}{A_k^2},$$

where A_k is the received amplitude of user k in the multiuser system.

An alternative definition is given by

$$\eta_k = \sup\left\{0 \leq r \leq 1\colon \lim_{\sigma\to 0}\frac{P_k(\sigma)}{Q\left(\frac{\sqrt{r}A_k}{\sigma}\right)} = 0\right\} \tag{19}$$

$$= \frac{2}{A_k^2}\lim_{\sigma\to 0}\log\frac{1}{P_k(\sigma)}. \tag{20}$$

This latter definition can be interpreted as one that compares the decay rate of the probability of error of a multiuser detector as a function of the SNR to that of the corresponding optimal single user system. If the probability of error does not go to zero with increasing SNR, then the asymptotic multiuser efficiency will be equal to 0. A unit asymptotic efficiency ($\eta_k = 1$) implies that asymptotically as the additive Gaussian noise vanishes, the multiuser detector being investigated for user k achieves performance that is commensurate with an optimal detector performing in a single-user system.

Due to the limitations of the conventional receiver in the presence of strong interferers, there is an interest to examine the sensitivity of a receiver to the received powers of interfering users. The near–far resistance (Lupas and Verdú, 1990) is the worst-case asymptotic multiuser efficiency over the amplitudes of the interfering users:

$$\bar{\eta}_k = \inf_{A_j>0,\, j\neq k}\eta_k. \tag{21}$$

We note this is a rather conservative measure of performance, but it does quantify a receiver's *near–far resistance*.

Prior to summarizing the values of these performance measures for several key receivers, we make a few observations specific to linear receivers, that is receivers of the form $\hat{b}_k = \mathrm{sgn}(\mathbf{c}_k^H\mathbf{y})$. First, we note that we can write the

[5] $Q(x) = \int_x^\infty 1/\sqrt{2\pi}\exp\{-v^2/2\}dv$.

closed form expression for the probability of error for this scenario. This is due to the fact that the soft statistic $\mathbf{c}_k^H \mathbf{r}$ has the distribution of a Gaussian mixture. First we note that

$$\mathbf{c}_k^H \mathbf{y} = \sum_{j=1}^{K} A_j b_j \mathbf{c}_k^H \rho_j + \mathbf{c}_k^H \mathbf{n}; \tag{22}$$

recall that ρ_j is the jth row (or column) of the cross-correlation matrix $\mathbf{R}$,

$$P_k^{\mathbf{c}_k}(\sigma) = \frac{1}{2^{K-1}} \times \sum_{b_j=\pm 1, j\neq k} Q\left(\frac{A_k \mathbf{c}_k^H \rho_k + \sum_{j\neq k} A_j b_j \mathbf{c}_k^H \rho_j}{\sigma\sqrt{\mathbf{c}_k^H R \mathbf{c}_k}}\right). \tag{23}$$

We note that while one can explicitly calculate, in closed form, the probability of detection error for linear receivers, the calculation of the performance is exponential in the number of active users. Thus, as the number of users increases, this calculation becomes computationally expensive. The probability of error of a linear receiver, in the high SNR region, is dominated by the Q-function with the smallest argument. Thus, the asymptotic multiuser efficiency of the receiver $\mathbf{c}_k$ is given by

$$\eta_k^{\mathbf{c}_k}(\sigma) = \frac{\max^2\left(0, \mathbf{c}_k^H \rho_k - \sum_{j\neq k} \frac{A_j}{A_k} \left|\mathbf{c}_k^H \rho_j\right|\right)}{\mathbf{c}_k^H R \mathbf{c}_k}. \tag{24}$$

We now specialize these results for several key linear receivers.

A. Conventional Receiver

As the conventional receiver is a linear receiver, the desired asymptotic multiuser efficiency is computed by substituting $\mathbf{c}_k = \mathbf{s}_k$

$$\eta_1^c = \max^2\left(0, 1 - \sum_{j\neq k} \frac{A_j}{A_k} |\rho_{jk}|\right) \tag{25}$$

$$\bar{\eta}_1^c = \begin{cases} 1 & \text{if } \rho_{jk} = 0 \ j \neq k \\ 0 & \text{else} \end{cases}. \tag{26}$$

We note that the ineffectiveness of the conventional receiver is clearly observed by the zero near–far resistance. There always exists a permutation of received amplitudes such that the conventional receiver will offer the worst-case probability of error of $\frac{1}{2}$. The exception to this scenario is when mutually orthogonal spreading waveforms are employed and thus $\mathbf{R} = \mathbf{I}_K$. Thus as previously observed, the conventional receiver is near–far limited.

B. Decorrelating Detector

If we let $\underline{\rho^{-1}}_k$ denote the kth row of the inverse correlation matrix $\mathbf{R}^{-1}$ and $\underline{\rho}$ denote the kth row of the correlation matrix, then it is clear that $\underline{\rho^{-1}}^T_k \underline{\rho}_j = 0 \quad \forall j \neq k$. Thus, the resultant performance measures for the decorrelator are given by

$$\eta_1^d = \frac{1}{R_{11}^{-1}} \tag{27}$$

$$\bar{\eta}_1^d = \frac{1}{R_{11}^{-1}}. \tag{28}$$

The decorrelating detector offers performance that is independent of the received amplitudes of the interferers. This occurs since the receiver zero-forces the interference. This total suppression comes at the expense of reducing the signal content due to the user of interest. Thus, there are scenarios in the low SNR regime where the conventional receiver offers better performance than that of the decorrelating detector.

C. Minimum Mean-Squared Error Receiver

The asymptotic multiuser efficiency of the MMSE receiver can be computed using the preceding formulae. The near–far resistance can be shown to be equivalent to that of the decorrelating detector. Thus, while the MMSE receiver offers a lower probability of error in the low SNR region in comparison to the decorrelating detector, as the amplitudes of the users' increase it achieves the same performance as the decorrelating detector

$$\bar{\eta}_1^m = \frac{1}{R_{11}^{-1}}. \tag{29}$$

D. Optimal Detector

Due to the fact that the jointly optimal detector and the individually optimal detector converge to each other as $\sigma \to 0$, the two detectors achieve the same asymptotic multiuser efficiency and thus, the same near–far resistance

$$\eta_1^o = \frac{\min_{\epsilon\in\{0,\pm 1\}^K, \epsilon_1 = 1} \epsilon^T \mathbf{A}\mathbf{R}\mathbf{A}\epsilon}{A_1^2} \tag{30}$$

$$\bar{\eta}_1^o = \frac{1}{R_{11}^{-1}}. \tag{31}$$

The numerator of the asymptotic multiuser efficiency expression above describes the minimum distance between the true, noiseless, received signal and the possible error event. Thus, the asymptotic multiuser efficiency is analogous to the normalized minimum distance used to describe optimal performance in a single-user inter-symbol

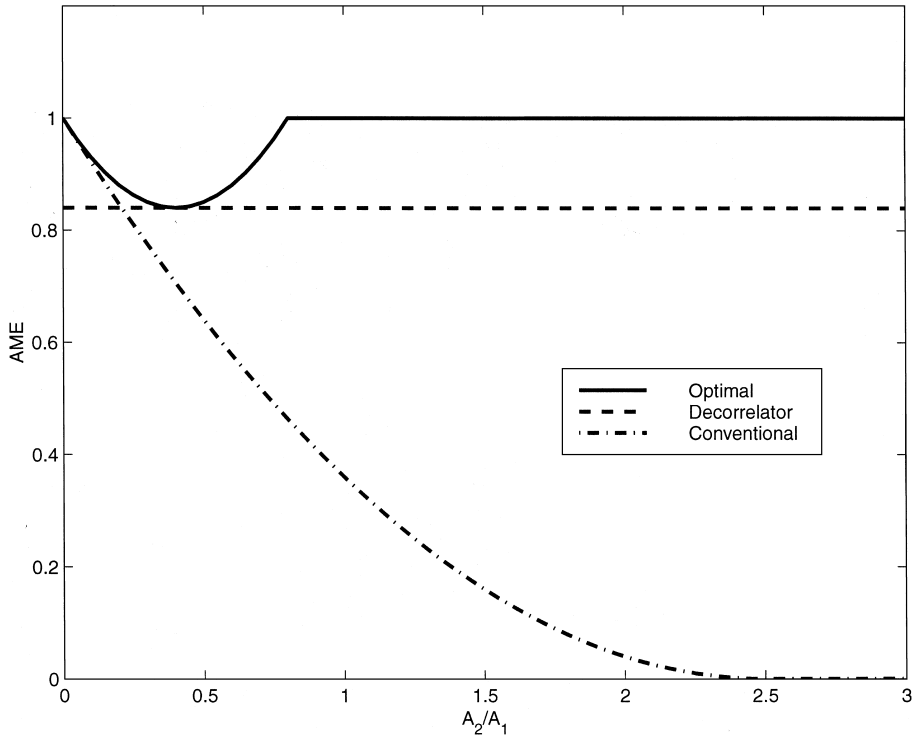


FIGURE 3 Asymptotic multiuser efficiency for several receivers; two-user case, $\rho = 0.4$.

interference channel (Proakis, 1989). We note that the decorrelating detector and the MMSE detector achieve optimal near–far resistance. The decorrelator is able to achieve this despite the fact that its construction is independent of the knowledge of the amplitudes of the active users. Both the optimal detector and the MMSE detector require this knowledge.

The asymptotic multiuser efficiency (AME) for a two-user system with $\rho = 0.4$ is shown in Fig. 3 for the jointly optimal detector, the decorrelator, and the conventional receiver. As predicted, the decorrelator has constant AME as a function of the near–far ratio, the conventional receiver's AME goes to zero, and the optimal detector achieves near unity AME for most near–far ratios.

VII. WIRELESS CHANNEL MODELS

Signals transmitted over wireless channels experience distortion due to a variety of phenomena; such as mobility and the effects of surrounding objects. The effect of reflection, scattering, and diffraction of transmitted signals by objects such as buildings, vehicles or fauna is termed *multipath*. An example of this effect occurs when trying to speak in a relatively small chamber with hard walls—there is a significant echo effect which hinders conversation. An abstract rendering of this phenomena is seen in Fig. 4.

As signals are received from different directions and with different propagation effects, the composite received signal is the sum of signals with different attenuations and delays. Considering the typical carrier frequencies employed in wireless communications, even small delays can result in large phase changes. The resulting signal can thus experience large variations in magnitude; this effect is called *fading*. Two other sources of attenuation are *path loss* and *shadowing*. Due to propagation losses,

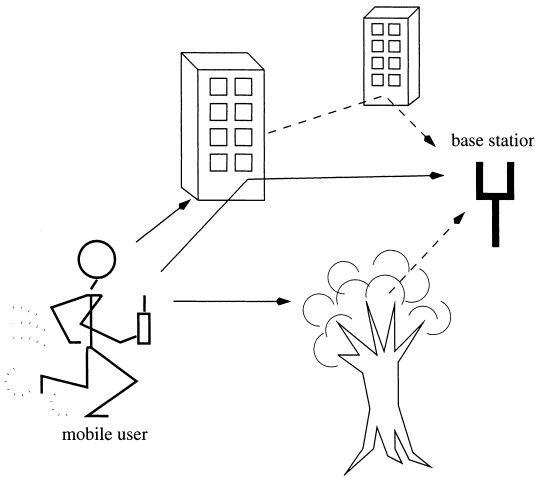


FIGURE 4 Multipath propagation.

the received power of a signal decays as the transmitter moves further away from the receiver; this effect is path loss. Typically path loss is proportional to $\frac{1}{d^\alpha}$ where d is the distance between the transmitter and receiver and α is the decay exponent. The effect of slow variations (relative to fading) of the mean of the envelope of the received signal is deemed shadowing; this phenomenon is induced by effects over distances corresponding to tens of the wavelength of the carrier frequency of the transmitted signal.

For a large number of scenarios, the wireless propagation channel can be accurately modeled as a linear, time-varying filter, $h(t;\tau)$ which is a wide-sense stationary complex random process with respect to t. The received signal is distorted in both time and frequency. The resultant spread of the signal in the frequency domain is due to the time-variation of the channel and the parameter which quantifies this effect is called the *Doppler spread.* Analogously, the signal also experiences time spreading due to the multipath; the corresponding parameter of interest is the *multipath delay spread.* These two values succinctly measure the overall distortion and can be best described by considering the Fourier transform of the channel with respect to t and τ, that is $\mathcal{H}(\lambda;\omega) = \mathcal{F}(h(t;\tau))$. Averaging this Fourier transform over the frequency variable ω, yields the *power spectrum* at time delay λ,

$$\mathcal{H}_c(\lambda) = \int_{-\infty}^{\infty} \mathcal{H}(\lambda, \omega)\, d\omega.$$

The multipath delay spread T_m is the largest value of λ for which $\mathcal{H}_c(\lambda)$ equals zero; it is a measure of the support of the function $\mathcal{H}_c(\lambda)$. As a measure of the time-spread, it indicates the amount of intersymbol interference present in the received signal. The inverse of the multipath delay spread is called the *channel coherence bandwidth*, $B_c = 1/T_m$. The parameter B_c describes the bandwidth of frequencies which experience correlated fading. Thus if $B_c = 200\,\text{Hz}$, we note that if the following three signals were transmitted,

$$s_1(t) = \cos(100\pi t)$$
$$s_2(t) = \cos(200\pi t)$$
$$s_3(t) = \cos(800\pi t)$$

then, $s_1(t)$ and $s_2(t)$ would experience correlated distortion while $s_1(t)$ and $s_3(t)$ would experience statistically independent channels.

The distortion induced by time-variation can be evaluated by considering the *Doppler power spectrum*,

$$\mathcal{H}_D(\omega) = \int_{-\infty}^{\infty} \mathcal{H}(\lambda, \omega)\, d\lambda$$

The *Doppler bandwidth* B_d is a measure of the support of the Doppler power spectrum and corresponds to the largest frequency such that $\mathcal{H}_D(\omega)$ equals zero. The inverse of the Doppler bandwidth is the coherence time of the channel, i.e. $T_{coh} = 1/B_d$; this value is an indicator of how fast the channel is fading. The Doppler bandwidth, which corresponds to the largest possible Doppler shift[6] can be determined by the carrier frequency and the velocity of the mobile,

$$B_d = \frac{v f_c}{c}, \tag{32}$$

where v is the velocity of the mobile in meters per second, f_c is the carrier frequency in Hz, and c is the speed of light (3×10^8 m/s).

If $T_m B_d < 1$, we say that the channel is *underspread.* The underspread channel can be estimated relatively accurately. In contrast, if $T_m B_d > 1$, the channel is deemed *overspread* and it is challenging to obtain high fidelity estimates of the channel, exacerbating equalization.

The channel experienced by a transmitted signal is a function of the transmitted signal bandwidth; as a result, proper models to describe the channel are a function of this bandwidth. For example, consider the Fourier transform of the channel with respect to τ only. Denote this function as $h(t;\omega)$. If the signal bandwidth W is significantly less than the coherence bandwidth of the channel then each frequency component of the transmitted signal experiences a common channel. Thus we can model $h(t;\omega) = h(t;0)\ \forall\omega$ and for a transmitted signal $s(t)$, the received signal $r(t)$ can be described as

$$\begin{aligned} r(t) &= \int_{-\infty}^{\infty} h(t;\tau)s(t-\tau)d\tau + n(t) \\ &= \int_{-\infty}^{\infty} h(t;\omega)S(\omega)e^{j2\pi\omega t}d\omega + n(t) \\ &= h(t;0)\int_{-\infty}^{\infty} S(\omega)e^{j2\pi\omega t} + n(t) \\ &= h(t)s(t) + n(t), \end{aligned}$$

where $n(t)$ is an additive white Gaussian noise process and $h(t;0) = h(t)$. Thus, the experienced channel is a time-varying multiplicative scalar. Furthermore, $h(t)$ is a random process with respect to t. In this scenario, the effects of the channel are clear: multiplication in the time domain is equivalent to convolution in the frequency domain. Thus, the wireless channel spreads the received signal in the frequency domain, the amount being equal to $2B_d$, the width of the Doppler spectrum.

In spread-spectrum systems, the signal bandwidth typically exceeds the coherence bandwidth of the channel;

[6]Note that the Doppler shift is a function of the angle of incidence of the plane wave impinging on the receiver.

then bands of frequencies experience independent fading termed *frequency selective fading*. In such a situation, it can be shown that different multipath components can be resolved and the corresponding channel model is given by

$$h(t,\tau) = \sum_{l=0}^{L-1} h_l(t)\delta\left(\tau - \frac{l}{W}\right),$$

where W is the bandwidth of the transmitted signal and the multipath coefficients $h_l(t)$ are modeled as random processes that are independent over the path number. The received signal is the convolution of the multipath channel with the transmitted signal, thereby resulting in a time spread equivalent to delay spread of the channel.

To complete discussion of the channel model, we require statistical descriptions of the channel tap coefficients $h_l(t)$.[7] Two commonly employed probability density functions for the magnitude of $h_l(t)$ ($|h_l(t)|$) are the Rayleigh distribution and the Ricean distribution:

$$\text{Rayleigh:}\quad p(h) = h\exp\left(-\frac{h^2}{2}\right) \tag{33}$$

$$\text{Ricean:}\quad p(h) = h\exp\left(-\frac{h^2+d^2}{2}\right)I_0(hd). \tag{34}$$

The function $I_0(x)$ is the zero-order modified Bessel function of the first kind.[8] And d is deemed the *noncentrality* parameter. The Rayleigh amplitude distribution occurs when there is uniform and diffuse scattering (typical of macro-cells (Stüber, 1996)), while the Ricean distribution occurs when there is a line-of-sight between the transmitter and the receiver, resulting in a nonzero mean for the random variable $h_l(t)$ (typical of micro-cells). In both scenarios, the channel coefficient $h_l(t)$ is a complex Gaussian random process with independent real and imaginary components.

Finally, the autocorrelation function of the Gaussian processes are required. For simplicity, we look at the autocorrelation of the real part of the channel coefficient. As the real and imaginary parts are identically distributed and independent, this is sufficient.

Let $\alpha(t) = Re(h_l(t))$. Then,

$$\mathbf{E}\{\alpha(t)\alpha^*(t+\tau)\} = \frac{P}{2}J_0(2\pi B_d\tau), \tag{35}$$

where

$$P = \mathbf{E}\left\{|h_l(t)|^2\right\}. \tag{36}$$

The function $J_0(x)$ is the zero-order Bessel function of the first kind. Furthermore, the power spectral density of the baseband real channel coefficient is given by

$$S_{\alpha\alpha}(f) = \begin{cases} \frac{P}{4\pi B_d}\frac{1}{\sqrt{1-(f/B_d)^2}} & |f| \le B_d \\ 0 & \text{otherwise} \end{cases}. \tag{37}$$

This channel correlation function and associated power spectral density are derived under the assumption of independent multipath channel phases which are further assumed to be uniformly distributed.

Given the transmission bandwidth of spread-spectrum systems, a multipath model is the model of choice. In the sequel, we discuss the received signal model for this scenario and methods for modifying the previously discussed receivers for the multipath case.

A. Receivers for Fading Multipath Channels

In this section, we consider the design of receivers for more realistic channel environments. While the optimal approach for asynchronous communication is the detection of the entire frame of data (sequence detection), we first discuss a conceptually simple approach which performs detection using observation intervals corresponding to single bit durations (one-shot detection).

1. One-Shot Receivers

We begin by introducing the filtered and sampled multipath channel model and providing the resultant signal model. In addition to the assumption of a more realistic channel, we also consider possible asynchronism among the active users. However, we shall assume that the multipath channels can be accurately modeled as linear and **time-invariant** filters. Thus, we use the time-invariant version of Eq. (33). Furthermore, the bandwidth of the DS-CDMA signals are determined by the chipping rate, thus $W \approx 1/T_c$. The result is a finite impulse response filter whose taps are spaced one chip (T_c) apart (Proakis, 1989; Stüber, 1996).

The received baseband signal corresponding to a single symbol interval, which is coherently demodulated, chip-matched filtered and sampled at the chip-rate is now described as

$$\mathbf{r}(i) = \sum_{k=1}^{K}\sum_{l=1}^{L_k} A_k\left[h_{kl}(i)b_k(i)\mathbf{s}_{kl}^{+} + h_{kl}(i-1)b_k \times (i-1)\mathbf{s}_{kl}^{-}\right] + \mathbf{n}(i). \tag{38}$$

We note that this received vector is of dimension $N \times 1$. Given that these relatively short observations will be employed to demodulate a single bit, we call such receivers *one-shot* or *n-shot* receivers. The variable n refers to the duration of the observation in symbol durations. In the sequel, we will discuss *sequence* detection, or the detection

[7] Note that for the single path, flat-fading case, there is only a single tap and $L = 1$.

[8] $I_0(x) = 1/2\pi \int_0^{2\pi} \exp\{-x\cos\theta\}d\theta$.

of every symbol in the transmitted frame. In general, for asynchronous, multipath systems, such short observations lead to degraded performance. We note, however, that the signal descriptions provided are easily generalized to the consideration of received signal vectors which correspond to multiple symbol intervals. The complex coefficients $h_{kl}(i)$ correspond to the multipath coefficients. The number of multipaths for user k is denoted by L_k. The partial spreading vectors $\mathbf{s}_{kl}^{+}$ and $\mathbf{s}_{kl}^{-}$ correspond to truncated and shifted versions of the spreading codes. For example, if we consider a two-user system, with $L_1 = 2$ and $L_2 = 1$, employing spreading codes of length $N = 7$, a possible realization of the partial spreading codes would be

$$\begin{aligned}
\mathbf{s}_{11}^{+} &= [\ 0 \quad 0 \quad \mathbf{s}_1(1) \quad \mathbf{s}_1(2) \quad \mathbf{s}_1(3) \quad \mathbf{s}_1(4) \quad \mathbf{s}_1(5)] \\
\mathbf{s}_{11}^{-} &= [\mathbf{s}_1(6) \quad \mathbf{s}_1(7) \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ ] \\
\mathbf{s}_{21}^{+} &= [\ 0 \quad 0 \quad 0 \quad 0 \quad \mathbf{s}_2(1) \quad \mathbf{s}_2(2) \quad \mathbf{s}_2(3)] \\
\mathbf{s}_{22}^{+} &= [\ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad \mathbf{s}_2(1) \quad \mathbf{s}_2(2)] \\
\mathbf{s}_{21}^{-} &= [\mathbf{s}_2(4) \quad \mathbf{s}_2(5) \quad \mathbf{s}_2(6) \quad \mathbf{s}_2(7) \quad 0 \quad 0 \quad 0\ ] \\
\mathbf{s}_{22}^{-} &= [\mathbf{s}_2(3) \quad \mathbf{s}_2(4) \quad \mathbf{s}_2(5) \quad \mathbf{s}_2(6) \quad \mathbf{s}_2(7) \quad 0 \quad 0\ ]
\end{aligned} \tag{39}$$

Here we note that user 1 has a delay of $\tau_1 = 2$ chips with respect to the receiver clock, while user 2 has a delay of $\tau_2 = 4$ chips. We can denote the discrete baseband equivalent representation of the multipath channel for user k as the complex vector $\mathbf{h}_k = [h_{k1}, h_{k2}, \ldots h_{kL_k}]^T$. The *effective spreading code* is the convolution of the transmitted spreading sequence with this channel vector and is denoted by $\bar{\mathbf{s}}_k = \mathbf{h}_k \star \mathbf{s}_k$, where $\star$ denotes convolution. Note that $\bar{\mathbf{s}}_k$ is of length $N + L_k - 1$. If we let

$$\bar{\mathbf{s}}_k^{+} = \sum_{l=1}^{L_k} h_{kl}\mathbf{s}_{kl}^{+} \tag{40}$$

$$\bar{\mathbf{s}}_k^{-} = \sum_{l=1}^{L_k} h_{kl}\mathbf{s}_{kl}^{-} \tag{41}$$

and

$$\bar{\mathbf{S}}^{+} = \left[\bar{\mathbf{s}}_1^{+}, \bar{\mathbf{s}}_2^{+}, \ldots \bar{\mathbf{s}}_K^{+}\right] \tag{42}$$

$$\bar{\mathbf{S}}^{-} = \left[\bar{\mathbf{s}}_1^{-}, \bar{\mathbf{s}}_2^{-}, \ldots \bar{\mathbf{s}}_K^{-}\right], \tag{43}$$

then we can rewrite the received signal vector as

$$\mathbf{r}(i) = \bar{\mathbf{S}}^{+}\mathbf{A}\mathbf{b}(i) + \bar{\mathbf{S}}^{-}\mathbf{A}\mathbf{b}(i-1) + \mathbf{n}(i). \tag{44}$$

From this description it is clear to see that the K user asynchronous multiuser system can be considered a $2K$ user synchronous system if observations corresponding to N chips (one symbol interval) are employed. We also observe the following relationship between $\bar{\mathbf{s}}_k$ and the vectors $\bar{\mathbf{s}}_k^{+}$, $\bar{\mathbf{s}}_k^{-}$. Let τ_k be the delay of user k in integer multiples of a chip then

$$\left[\bar{\mathbf{s}}_k^{+}, \bar{\mathbf{s}}_k^{-}\right] = \left[\underbrace{0, 0, \ldots 0}_{1\times\tau_k},\ \bar{\mathbf{s}}_k,\ \underbrace{0, 0, \ldots, 0}_{(N-\tau_k-L_k+1)\times 1}\right]. \tag{45}$$

We note that if the channel of the desired user is completely known, i.e., we have knowledge of $\mathbf{h}_k$ and τ_k, then the previous algorithms can be applied directly where observations of length $N + L_k - 1$ are collected and the desired user's spreading code, $\mathbf{s}_k$, is replaced by the effective spreading code, $\bar{\mathbf{s}}_k$. If information about the interfering users is necessary, the interfering users are modeled as $2(K-1)$ synchronous users with spreading waveforms $\bar{\mathbf{s}}_k^{+}$ and $\bar{\mathbf{s}}_k^{-}$. Thus receivers can be constructed with the view that each user is viewed as a *single* user with spreading code $\bar{\mathbf{s}}_k$. We note that an alternative view is possible which can obviate the need for the channel vector $\mathbf{h}_k$, but necessitates knowledge of the channel length L_k and the relative delay τ_k. In this approach, each user is considered to be L_k users with a spreading code that is a shifted version of the other spreading codes. Then, receivers can be designed for each path. The final decision is made by combining the soft decisions from each of the L_k adaptive receivers. The choice of combining coefficients remains an open question. A typical approach is to consider *equal gain combining* as in Barbosa and Miller (1998). This approach can be applied to the bulk of the receivers considered in Section V. The difference between these two approaches is depicted in Fig. 5.

Examples of receiver designs based on the ideas expressed in this section are found in Huang and Papadias (1998) and Wang and Poor (1999).

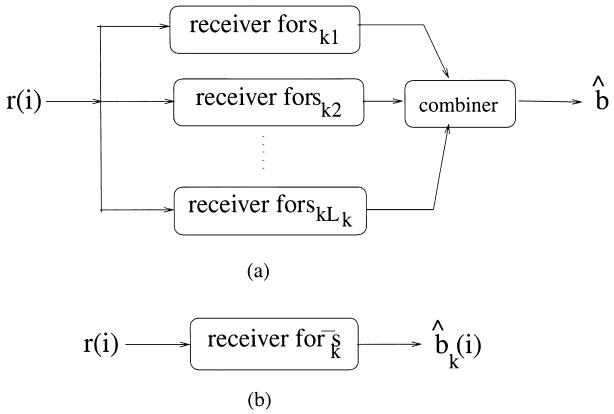


FIGURE 5 (a) A virtual user approach to multipath multiuser detection, where one receiver is constructed for each path of the channel. (b) An effective spreading code approach where the spreading waveform $\bar{\mathbf{s}}_k$ is the convolution of the channel with the spreading code.

2. Sequence Detectors

In the previous section, detectors for a single bit were discussed for realistic channel conditions. In the current section, we shall focus on static channels and asynchronous communication to discuss sequence detectors. Given a frame of transmitted $2M+1$ bits: $\tilde{\mathbf{b}}$ is a data vector of length $(2M+1)K$, where $\tilde{\mathbf{b}}(k+iK) = b_k(i)$ for $k = 1, \ldots, K$ and $i = -M, \ldots M$. The received signal in continuous time is given by

$$r(t) = \sum_{i=-M}^{M} \sum_{k=1}^{K} A_k b_k(i) s_k(t - iT - \tau_k) + n(t),$$

where we recall that τ_k is the delay for user k. We assume that the users are ordered such that $\tau_1 \leq \tau_2, \ldots \leq \tau_k$. If we define the following virtual user spreading waveform $\tilde{s}_k(t)$,

$$\tilde{s}_{k+iK}(t) = s_k(t - iT - \tau_k),$$

then a set of $(2M+1)K$ matched filter outputs, $\tilde{\mathbf{y}}$ are formed by

$$\tilde{\mathbf{y}}_j = \int_{-\infty}^{\infty} \tilde{s}_j(t) r(t)\, dt.$$

The resulting matched filter outputs can be expressed as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{R}}\tilde{\mathbf{A}}\tilde{\mathbf{b}} + \tilde{\mathbf{n}},$$

where

$$\tilde{\mathbf{R}}(i, j) = \int_{-\infty}^{\infty} \tilde{s}_i(t)\tilde{s}_j(t)\, dt$$

$$\tilde{\mathbf{A}}(k+iK, k+iK) = A_k$$

$$\tilde{\mathbf{n}}_j = \int_{-\infty}^{\infty} \tilde{s}_j(t) n(t)\, dt.$$

Note that the diagonal amplitude matrix $\tilde{\mathbf{A}}$ is constructed by taking the diagonal of the concatenation of the vector $[A_1, \ldots A_K]$ $2M+1$ times. The jointly optimal detector selects the vector $\hat{\tilde{\mathbf{b}}}$ such that,

$$\hat{\tilde{\mathbf{b}}} = \arg \max_{\tilde{\mathbf{b}} \in \{\pm 1\}^{(2M+1)K}} \Omega(\tilde{\mathbf{b}}) = 2\tilde{\mathbf{b}}^T \tilde{\mathbf{A}}\tilde{\mathbf{y}} - \tilde{\mathbf{b}}^T \tilde{\mathbf{A}}\tilde{\mathbf{R}}\tilde{\mathbf{A}}\tilde{\mathbf{b}}.$$

The complexity of the preceding optimization is $O(2^{2M+1)K})$; this can be prohibitively large for reasonable frame sizes M. However, the complexity can be reduced by recognizing that the matrix $\tilde{\mathbf{R}}$ is sparse. In fact, $\Omega(\tilde{\mathbf{b}})$ can be written as

$$\Omega(\tilde{\mathbf{b}}) = \sum_{j=1-MK}^{MK+K} \omega_j(\mathbf{x}_j, b_j)$$

where

$$= \mathbf{x}_{j+1} = [x_{j+1}(1), \ldots, x_{j+1}(K-1)]$$
$$= [x_j(2), \ldots, x_j(K-1), b_j].$$

The system is initialized with $\mathbf{x}_{1-MK} = [0, \ldots 0]$. The vector $\mathbf{x}$ can be considered to be the state of a shift-register with $K-1$ registers. Thus a Viterbi algorithm can be employed to determine the sequence estimate of $\tilde{\mathbf{b}}$ and the demodulator algorithm can be described by a trellis with 2^{K-1} states. The complexity of this implementation is $O(2^K)$.

Sequence detectors based on zero-forcing (decorrelating) and minimum-mean squared error criteria can be constructed by considering the Z-transform of the matched filter output corresponding to the symbol timing i $\mathbf{y}(i) = [\tilde{\mathbf{y}}_{1+iK}, \tilde{\mathbf{y}}_{2+iK}, \ldots \tilde{\mathbf{y}}_{(i+1)K}]^T$. It can be shown that the Z-transform of these matched filtered outputs can be represented as

$$\mathbf{Y}(z) = \mathbf{S}(z)\mathbf{A}\mathbf{B}(z) + \mathbf{N}(z),$$

where $\mathbf{S}(z)$ is the channel transfer function which incorporates the effects of multiple access interference, $\mathbf{A}$ is the previously defined $K \times K$ diagonal amplitude matrix, and $\mathbf{B}(z)$ is the Z-transform of the data stream. The linear transformations of the decorrelating detector and the minimum-mean squared error receiver in the Z-domain are

- Decorrelating sequence detector

$$\mathbf{M}(z) = \mathbf{S}(z)^{-1}.$$

- Minimum mean squared error sequence detector

$$\mathbf{M}(z) = (\mathbf{S}(z) + \sigma^2 \mathbf{A}^{-2})^{-1}.$$

VIII. ADDITIONAL RECEIVER STRUCTURES

In this section, a host of receiver algorithms for a variety of specific communication environments are considered. Both training-based and blind adaptive algorithms are described; reduced rank methods are explored; receiver algorithms for multimedia traffic are considered; and methods for decoding coded multiuser transmissions are investigated.

A. Adaptive Algorithms

The objective of this section is to introduce some methods for the adaptive demodulation of data in DS-CDMA systems. Adaptive algorithms are instrumental for providing consistent performance in unknown or time-varying

environments. Adaptive methods can implicitly learn unknown parameters of a system or can be used to track these parameters as they change over time. These characteristics of adaptive methods make them particularly suitable for wireless communications systems. In contrast to communication environments with relatively fixed characteristics, as is found in the wired environment of classical telephony, the wireless communication channel as noted before is time-varying (Stüber, 1996). This time-varying nature is due to the mobility of the transmitter as well as the dynamic user environment.

Two types of adaptive algorithms will be considered. In the first case, a *training signal* will be used to update the adaptive algorithm. This data signal is known at both the transmitter and receiver. In contrast, in *blind* adaptive algorithms, there is no training signal, cost functions based on the statistics of the signal are formed and updated.

1. Training-Based Adaptive Algorithms

For the adaptive least-mean squares (LMS) algorithm, the cost function is the mean-squared error:

$$J(\mathbf{c}) = \mathbf{E}\left\{\left|b_1(i) - \mathbf{c}^H\mathbf{r}(i)\right|^2\right\}. \tag{46}$$

The parameter vector is the desired linear receiver, which operates as follows,

$$\hat{b}_1(i) = \operatorname{sgn}\left(\mathbf{c}_{\mathrm{LMS}}(i)^H\mathbf{r}(i)\right). \tag{47}$$

The estimate of the parameter vector at time i ($\mathbf{c}(i)$) is obtained by employing the method of steepest descent. Thus,

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu\nabla J(\mathbf{c}(i-1)) \tag{48}$$

$$= \mathbf{c}(i-1) - \mu\mathbf{E}\left\{\mathbf{r}(i)(b_1(i-1) - \mathbf{c}(i-1)^H\mathbf{r}(i))\right\} \tag{49}$$

$$\approx \mathbf{c}(i-1) - \mu\mathbf{r}(i)\left(b_1(i-1) - \mathbf{c}(i-1)^H\mathbf{r}(i)\right). \tag{50}$$

The recursive least squares algorithm differs from the adaptive Least Mean Squares (LMS) algorithm just derived in that the cost function is a deterministic one. For an observation record corresponding to M symbols, it is desired to minimize the following metric,

$$J(\mathbf{c}) = \sum_{i=1}^{M} \lambda^{M-i}|e(i)|^2, \tag{51}$$

where λ is deemed the *forgetting factor* and is assumed to be $0 < \lambda < 1$. With this form of weighting, recent observations are given more weight than past observations. As before, the desired parameter vector is a linear receiver, thus

$$\hat{b}_1(i) = \operatorname{sgn}\left(\mathbf{c}_{\mathrm{RLS}}(i)^H\mathbf{r}(i)\right). \tag{52}$$

The resultant algorithm, which exploits the matrix inversion lemma (to avoid a computationally expensive direct matrix inverse) is given as follows:

$$\mathbf{k}(i) = \frac{\lambda^{-1}\mathbf{P}(i-1)\mathbf{r}(i)}{1 + \lambda^{-1}\mathbf{r}(i)^H\mathbf{P}(i-1)\mathbf{r}(i)} \tag{53}$$

$$\zeta(i) = b_1(i) - \mathbf{c}(i-1)^H\mathbf{r}(i) \tag{54}$$

$$\mathbf{c}(i) = \mathbf{c}(i-1) + \mathbf{k}(i)\zeta(i) \tag{55}$$

$$\mathbf{P}(i) = \lambda^{-1}\mathbf{P}(i-1) - \lambda^{-1}\mathbf{k}(i)\mathbf{r}(i)^H\mathbf{P}(i-1). \tag{56}$$

The typical initialization of the algorithm is with $\mathbf{P}(0) = \delta^{-1}\mathbf{I}_N$ and $\mathbf{c}(0) = \mathbf{0}_N$, where δ is a small positive constant. In the multiuser receiver context, we can also initialize the weight vector to be $\mathbf{c}(0) = \mathbf{s}_1$, thus, the receiver is initialized as the conventional matched filter receiver. The vector $\mathbf{k}(i)$ is called the *gain vector*. We also note that $\mathbf{P}(i)$ is the current estimate of the inverse of the weighted (each observation is weighted by λ) data correlation matrix.

For the general case (linear estimation of a scalar process), one can show that the RLS algorithm has a rate of convergence which is far superior to that of the LMS algorithm (Haykin, 1996), however, this comes as the expense of greater computational complexity. Theoretically, the RLS algorithm produces zero excess mean-squared error. This result is dependent on the presence of a statistically stationary environment, which is not typically experienced in a wireless communications channel. Finally, the convergence of the RLS algorithm is not dependent on the eigenvalue spread of the data correlation matrix. In contrast, the LMS algorithm's convergence speed is dependent on this eigenvalue spread. The implication for DS-CDMA systems is potential slow convergence as the eigenvalue spread can be quite large in a system where there is a great disparity in received powers amongst the users. Thus, if tight power control is not in place, the LMS algorithm will experience slow convergence. However, for many scenarios, this improved convergence speed for RLS is, in fact, not observed (Woodward and Vucetic, 1998).

2. Blind Adaptive Algorithms

Next, adaptive algorithms which do not require a training sequence are described. We first consider an algorithm which employs the data on a sample by sample basis and subsequently describe an algorithm which requires the estimation of second order statistics. Such methods are titled *subspace*-based techniques as they use the estimates of the second order statistics to estimate the subpsace spanned by the spreading waveforms of the active users as well as the corresponding noise subspace, which is the

orthogonal complement of the signal subspace. Properties of well-chosen spreading waveforms can then be exploited to estimate an equalizer for the desired user based on the noise or signal subspace estimate.

A blind adaptive multiuser receiver was suggested in Honig (1995). The cost function for this receiver is the *mean output energy* (MOE),

$$J(\mathbf{c}) = \mathbf{E}\left\{\left|\mathbf{c}^H\mathbf{r}(i)\right|^2\right\} \tag{57}$$

$$= \mathbf{c}^H\mathbf{C}\mathbf{c} = \mathbf{E}\left\{\left|(\mathbf{x}_1 + \mathbf{s}_1)^H\mathbf{r}(i)\right|^2\right\}, \tag{58}$$

where

$$\mathbf{c} = \mathbf{x}_1 + \mathbf{s}_1 \tag{59}$$

and

$$\mathbf{x}_1^H\mathbf{s}_1 = 0. \tag{60}$$

It is straightforward to show that the vector $\mathbf{x}_1$ which minimizes the mean output energy also minimizes the mean-squared error. That is $\mathbf{c}_{\text{MMSE}} = \mathbf{s}_1 + \mathbf{x}_{1_{\text{MOE}}}$. Honig *et al.* (1995) term the decomposition $\mathbf{c}_1 = \mathbf{x}_1 + \mathbf{s}_1$ a *canonical* decomposition of the desired receiver vector into a signal which lies along the desired users' signal subspace $\mathbf{s}_1$ plus a signal which is orthogonal to the desired user's signal subspace ($\mathbf{x}_1$). If we ignore the orthogonality constraint, then the receiver vector update algorithm is given in the following:

$$\mathbf{c}(i) = \mathbf{c}(i-1) - \mu\Delta J(\mathbf{c}(i-1)) \tag{61}$$

$$= \mathbf{c}(i-1) - \mu\mathbf{E}\{\mathbf{r}(i)^H\mathbf{c}(i-1)\mathbf{r}(i)\} \tag{62}$$

$$\approx \mathbf{c}(i-1) - \mu\mathbf{r}(i)^H\mathbf{c}(i-1)\mathbf{r}(i). \tag{63}$$

However, we wish to ensure that the vector $\mathbf{x}_1$ remains orthogonal to $\mathbf{s}_1$. Thus, we use the projected approximate gradient rather than the true gradient to update the orthogonal vector $\mathbf{x}_1$. That is,

$$\mathbf{r}(i) - \left(\mathbf{r}(i)^H\mathbf{s}_1\right)\mathbf{s}_1 \tag{64}$$

is orthogonal to $\mathbf{s}_1$. The orthogonal signal update is then

$$\mathbf{x}_1(i) = \mathbf{x}_1(i-1) - \mu\left(\mathbf{r}(i)^H(\mathbf{s}_1 + \mathbf{x}_1(i-1))\right)$$
$$\times\left(\mathbf{r}(i) - \left(\mathbf{r}(i)^H\mathbf{s}_1\right)\mathbf{s}_1\right). \tag{65}$$

We next consider a blind algorithm that relies on the estimation of second order statistics rather than updating a receiver recursively. The subspace based estimation methods of Wang and Poor (1998a,b) exploit the properties of the subspace decomposition of the data correlation matrix. The data correlation matrix can be represented as follows:

$$\mathbf{C} = \mathbf{E}\{\mathbf{r}\mathbf{r}^T\} = \sigma^2\mathbf{I} + \sum_{k=1}^{K} A_k^2\mathbf{s}_k\mathbf{s}_k^T \tag{66}$$

$$= \sigma^2\mathbf{I} + \mathbf{S}\mathbf{A}^2\mathbf{S}^T. \tag{67}$$

We consider a spectral decompostion of the contribution due to the users only,

$$\mathbf{S}\mathbf{A}^2\mathbf{S}^T = \mathbf{V}\Lambda\mathbf{V}^T \tag{68}$$

where

$$\mathbf{V}^T\mathbf{V} = \mathbf{I} \tag{69}$$

and

$$\Lambda = \text{diag}\left[\lambda_1, \lambda_2, \ldots, \lambda_K\right]. \tag{70}$$

Note that the span of the columns of $\mathbf{V}$ is equivalent to the span of the columns of $\mathbf{S}$, this space is called the *signal subspace.* The orthogonal complement of the signal subspace is denoted the *noise subspace.* A related decomposition for the data correlation matrix can be found by augmenting the eigenvector matrix $\mathbf{V}$ such that

$$\mathbf{C} = \tilde{\mathbf{V}}\Gamma\tilde{\mathbf{V}}^T \tag{71}$$

where

$$\tilde{\mathbf{V}} = [\mathbf{V}, \mathbf{V}'] \tag{72}$$

$$\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \mathbf{I} \tag{73}$$

$$\Gamma = \text{diag}\left[\gamma_1, \gamma_2, \ldots, \gamma_K\right] \tag{74}$$

$$\gamma_i = \begin{cases} \lambda_i + \sigma^2 & i = 1, \ldots, K \\ \sigma^2 & i = K+1, \ldots, N. \end{cases} \tag{75}$$

Using these definitions, it can be shown that the decorrelating detector and the MMSE detector can be represented in terms of the aforementioned eigenvectors and eigenvalues:

$$\mathbf{c}_k^{DD} = \mathbf{V}\,\text{diag}^{-1}\left[\lambda_1 - \sigma^2, \ldots, \lambda_K - \sigma^2\right]\mathbf{V}^T\mathbf{s}_k \tag{76}$$

$$\mathbf{c}_k^{MMSE} = \tilde{\mathbf{V}}\Gamma^{-1}\tilde{\mathbf{V}}^T\mathbf{s}_k. \tag{77}$$

In order to construct a blind implementation, a sample data correlation matrix estimate is formed. The necessary eigen decompositions are done using this estimated matrix. Note that the decorrelator is a function of the signal space only, while the MMSE receiver is a function of both the noise and the signal subspace. Examples of other subspace based algorithms include Liu and Xu (1995), Liu and Zoltowski (1997), Sorelius *et al.* (1998), and Tsatsanis and Giannakis (1997).

B. Reduced Rank Methods

Reduced rank methods have received recent attention for wireless communication systems. This is due to the fact that in time-varying channels, full-rank adaptive methods have an insufficiently fast convergence rate and often cannot track a time-varying channel effectively. If fewer tap coefficients have to be adapted, the convergence rate is often faster. To introduce the reduced rank methods, we

recall the full-rank fixed MMSE receiver for the desired user 1. This linear MMSE receiver is an $N \times 1$ vector $\mathbf{c}$ which minimizes the mean-squared error (MSE),

$$\mathbf{c}_{\text{MMSE}} = \arg\min_{\mathbf{c}} \text{MSE} = \arg\min_{\mathbf{c}} \mathbf{E}\left\{\left|b_1(i) - \mathbf{c}^H \mathbf{r}(i)\right|^2\right\} = \mathbf{C}^{-1}\mathbf{p}. \quad (78)$$

Recall that $\mathbf{C} = \mathbf{E}\{\mathbf{r}(i)\mathbf{r}(i)^H\}$ is the data cross-correlation matrix and the correlation vector, $\mathbf{p} = \mathbf{E}\{b_1(i)\mathbf{r}(i)\}$, is termed the *steering vector* (Miller *et al.*, 2000).

The adaptive algorithms (Miller *et al.*, 2000; Zhu *et al.*, 2000) can be used to estimate $\mathbf{c}_{\text{MMSE}}$, even in a time-varying channel. When N is large, convergence is slow. Reduced rank techniques reduce the number of taps to be adaptively tracked by projecting the received signal vector onto a lower-dimensional subspace. Let D be the resultant lower dimension, where $D < N$, the projection is

$$\tilde{\mathbf{r}}(i) = \mathbf{P}_D^H \mathbf{r}(i), \quad (79)$$

where $\mathbf{P}_D$ is the $N \times D$ projection matrix, and the D dimensional signal is denoted by a "tilde." The vector $\tilde{\mathbf{r}}(i)$ is then the input to a length D tap delay linear filter. When the MMSE criterion is applied, the optimum coefficients for the D dimensional space are given by

$$\tilde{\mathbf{c}}_{\text{MMSE}} = \tilde{\mathbf{C}}^{-1}\tilde{\mathbf{p}}, \quad \text{where } \tilde{\mathbf{C}} = \mathbf{P}_D^H \mathbf{C} \mathbf{P}_D, \tilde{\mathbf{p}} = \mathbf{P}_D^H \mathbf{p}. \quad (80)$$

1. Projection Matrix Selection

Herein, a number of methods for selecting the projection matrix $\mathbf{P}_D$ are considered.

The multistage Wiener filtering (MWF) algorithm for DS-CDMA was presented in Honig and Goldstein (2000). The resultant algorithm is a specialization of the work in Goldstein *et al.* (1998). The MWF projection matrix is given by

$$\mathbf{P}_D^{MW} = [\mathbf{g}_{MW,1} \quad \mathbf{g}_{MW,2} \quad \cdots \quad \mathbf{g}_{MW,D}] = \left[\mathbf{h}_1 \quad \mathbf{B}_1^H \mathbf{h}_2 \quad \cdots \quad \prod_{j=1}^{D-1} \mathbf{B}_j^H \mathbf{h}_D\right], \quad (81)$$

where $\mathbf{g}_{MW,j}$, $j = 1, \ldots, D$ are implicitly defined. The matrix $\mathbf{B}_j$ is an $(N-j) \times (N-j+1)$ blocking matrix, i.e., $\mathbf{B}_j \mathbf{h}_j = 0$. The vector $\mathbf{h}_j$ is the normalized correlation vector $E\{\tilde{d}_{j-1}(i)\mathbf{r}_{j-1}(i)\}$, where $\mathbf{r}_j(i) = \mathbf{B}_j \mathbf{r}_{j-1}(i)$ with $\mathbf{r}_0(i) = \mathbf{r}(i)$, and $\tilde{d}_j(i) = \mathbf{h}_i^H \mathbf{r}_{j-1}(i)$ with $\tilde{d}_0(i) = b_1(i)$ (Honig and Goldstein, 2000).

The projection matrix for the auxiliary vector filtering (AVF) algorithm is given by Pados *et al.* (1999),

$$\mathbf{P}_D^{AV} = [\mathbf{g}_{AV,1} \quad \mathbf{g}_{AV,2} \quad \cdots \quad \mathbf{g}_{AV,D}], \quad (82)$$

where $\mathbf{g}_{AV,1}$ is also the normalized correlation vector $E\{b_1(i)\mathbf{r}(i)\} = \mathbf{h}_1$, and $\mathbf{g}_{AV,j}$, $j = 2, \ldots, D$ are the auxiliary vectors, given by Pados *et al.* (1999)

$$\mathbf{g}_{AV,j+1} = \frac{\mathbf{C}\mathbf{g}_{AV,j}^{Eq} - \left(\mathbf{g}_{AV,1}^H \mathbf{C}\mathbf{g}_{AV,j}^{Eq}\right)\mathbf{g}_{AV,1} - \sum_{l=2}^{j}\left(\mathbf{g}_{AV,l}^H \mathbf{C}\mathbf{g}_{AV,j}^{Eq}\right)\mathbf{g}_{AV,l}}{\left\|\mathbf{C}\mathbf{g}_{AV,j}^{Eq} - \left(\mathbf{g}_{AV,1}^H \mathbf{C}\mathbf{g}_{AV,j}^{Eq}\right)\mathbf{g}_{AV,1} - \sum_{l=2}^{j}\left(\mathbf{g}_{AV,l}^H \mathbf{C}\mathbf{g}_{AV,j}^{Eq}\right)\mathbf{g}_{AV,l}\right\|}, \quad (83)$$

where $\mathbf{g}_{AV,j}^{Eq} = \mathbf{g}_{AV,1} - \sum_{l=2}^{j} w_l \mathbf{g}_{AV,j}$, w_l, $l = 2, \ldots, j$ are the optimized constants (Pados *et al.*, 1999) and $\|.\|$ is the vector norm. By construction, the $\mathbf{g}_{AV,j}$, $j = 1, \ldots, D$ are normalized orthogonal vectors.

Using the Cayley–Hamilton (CH) theorem, the following projection matrix (Moshavi *et al.*, 1996) was proposed,

$$\mathbf{P}_D^{CH} = [\mathbf{g}_{CH,1} \quad \mathbf{g}_{CH,2} \quad \cdots \quad \mathbf{g}_{CH,D}] = \left[\mathbf{h}_1 \quad \mathbf{C}\mathbf{h}_1 \quad \cdots \quad \mathbf{C}^{D-1}\mathbf{h}_1\right]. \quad (84)$$

The vector $\mathbf{h}_1$ is the one defined previously.

The projection matrix can also be selected by performing an eigendecomposition on $\mathbf{C}$, that is,

$$\mathbf{C} = \mathbf{V}\Lambda\mathbf{V}^H. \quad (85)$$

The matrix $\Lambda = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_N]$ contains the eigenvalues associated with the eigenvectors which are the columns of $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$. If the eigenvalues are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, the desired projection matrix is then $\mathbf{P}_D = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D]$. This method is titled the method of *principal components*. A related method, termed the *cross-spectral* reduced rank method, selects the D eigenvectors which result in a minimum mean-squared error estimate of the bit. This method requires knowledge of the desired user's spreading waveform, but offers improved performance over the principal components methods (Goldstein and Reed, 1997). Eigendecomposition methods, in general, are not attractive due to their high attendant computational complexity.

Finally, a very simple method of rank reduction is to employ partial despreading as proposed in Singh and Milstein (1999). The desired user's spreading code is decomposed

into multiple subvectors: $\mathbf{s}_1 = [\mathbf{s}_1^{(1)}, \mathbf{s}_1^{(2)}, \ldots, \mathbf{s}_1^{(D)}]$. The projection matrix is then constructed as

$$\mathbf{P}_D = \begin{bmatrix} \mathbf{s}_1^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{s}_1^{(2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{s}_1^{(D)} \end{bmatrix}. \tag{86}$$

Through a key simplification of the AVF algorithm, it can be shown that the AVF method is equivalent to the MWF algorithm (Chen *et al.*, 2000). The work in Chen *et al.* provides a simplified proof of the equivalence of the static MWF receiver and that based on the Cayley Hamilton expansion of the correlation matrix inverse (Moshavi *et al.*, 1996). This observation was previously made in Honig and Goldstein (2000).

C. Receivers for Multimedia Systems

With the discussion of future standards (Adachi *et al.*, 1998; Dahlman *et al.*, 1998; Okamura and Adachi, 1998), there has been heightened interest in wireless systems that offer multiple data rates in an integrated fashion. In such a system, users can consider transmitting at one of a class or possible data rates. The methods by which one can modify a data rate are varied. We describe a few here:

1. Multicode access (Fig. 6a). In this access method there exists a single constant-bit rate on all channels. High rate users are accommodated by multiplexing their data onto multiple signature sequences. This method offers the system construction advantage of being able to employ a single-rate receiver.
2. Variable sequence length (Fig. 6b). In this access method, different data rates are treated with spreading codes of different lengths while maintaining a common chipping rate.
3. Variable chip rate (Fig. 6c). Alternatively, one can keep the length of the signature sequences constant but vary the underlying chipping rate to accommodate varying user rates. Such a scheme has bandwidth-related implications as the various chipping rates will result in different bandwidth usages. It may be necessary to consider some frequency planning to employ this method of multi-rate communication. The differences in overall bandwidth will enable the receiver to view users as wider band and narrower band users. A multiple chip rated system has been considered for DS-CDMA communications (Wu and Geraniotis, 1994) again a conventional receiver was employed.

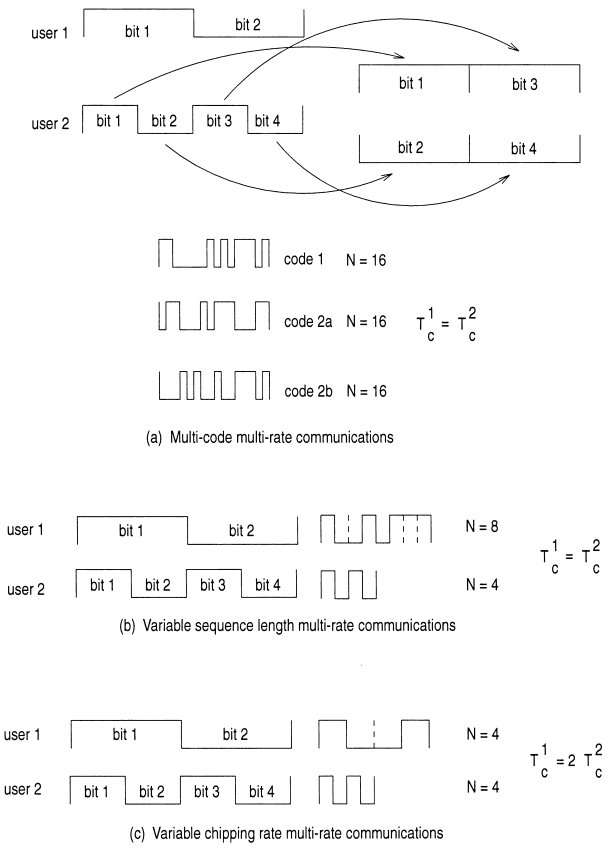


FIGURE 6 The various multi-rate multi-access schemes for DS-CDMA communication.

4. Variable modulation. This scheme is not specific to DS-CDMA; however, it is possible to modify the data rate of a user by adjusting the modulation scheme. For example, from switching from BPSK to QPSK, the data rate is doubled, however this will come at the expense of performance.

We note that single-rate receivers can be constructed for multicode CDMA systems; thus previously described receiver algorithms can be employed. Herein, we shall focus on systems where the symbol rates among users are different. Digital communication signals are, in general, cyclostationary. That is, given a received signal $r(t)$, is wide-sense cyclostationary with period T if

$$\mathbf{E}\{r(t)\} = \mathbf{E}\{r(t+T)\}$$

$$\mathbf{E}\{r(t_1)r^*(t_2)\} = \mathbf{E}\{r(t_1+T)r^*(t_2+T)\}.$$

For a digital communications signal, this period T is the symbol duration. In a multiuser environment where all users transmit at the same data rate, the period of cyclostationarity is also T; if multiple users have different symbol rates, then the received sum signal retains its cyclostationary nature. However, the period of cyclostationarity

is now the least common multiple of the individual symbol rates, that is $T^* = \mathrm{LCM}(T_1, T_2, \ldots, T_K)$. The key to proper receiver design in multirate multiuser systems is the adoption of a signal model which clearly considers the cyclostationary nature of the received signal. For simplicity of description, we consider a multirate system where the symbol durations are in powers of 2, that is $\mathcal{T} = \{T_1, T_2, \ldots, T_K\} = \{1, 2, 4, \ldots, 2^R\} T_{\min}$. We furthermore assume that the chip durations of each user are also multiple integers of each other, thus $T_c^{(k)} = \beta_k T_c^*$, where $T_c^{(k)}$ is the chip duration for user k, $T_c^* = \min_k T_c^{(k)}$ and $\beta_k \in Z^+$. Under the assumption of synchronrous communication, the coherently demodulated received signal is sampled at the chip-rate, $\frac{1}{T_c^*}$. Samples are collected such that signal vectors correspond to observations of duration T^*. Over this duration, each user can contribute multiple information bits; as such each user can be considered as multiple virtual users with zero-padded spreading vectors. We define the rate ratio for user k as $P_k = \frac{T^*}{T_k}$, P_k indicates the number of virtual users in the system over the duration T^*. The virtual user spreading code is defined to be

$$\mathbf{s}_k^{(p)}(j) = \begin{cases} \mathbf{s}_k\left(j - (p-1)\frac{T_k}{T_c^*}\right) & (p-1)\frac{T_k}{T_c^*} \le j < p\frac{T_k}{T_c^*}, \\ & p = 1, \ldots, P_k. \\ 0 & \text{else} \end{cases}$$

The received signal is specified by

$$\tilde{\mathbf{r}}(i) = \tilde{\mathbf{S}}\tilde{\mathbf{A}}\tilde{\mathbf{b}}(i) + \tilde{\mathbf{n}}(i), \tag{87}$$

where

$$\tilde{\mathbf{S}} = \left[\mathbf{s}_1^{(1)}, \ldots, \mathbf{s}_1^{(P_1)}, \mathbf{s}_2^{(1)}, \ldots, \mathbf{s}_2^{(P_2)}, \ldots, \mathbf{s}_K^{(P_K)}\right] \tag{88}$$

$$\tilde{\mathbf{A}} = \mathrm{diag}\left[\underbrace{A_1, \ldots, A_1}_{1\times P_1}, \ldots, \underbrace{A_K, \ldots, A_K}_{1\times P_K}\right] \tag{89}$$

$$\tilde{\mathbf{b}} = \left[b_1^{(1)}, b_1^{(2)}, \ldots, b_1^{(P_1)}, b_2^{(1)}, \ldots, b_2^{(P_2)}, \ldots, b_K^{(P_K)}\right]^T. \tag{90}$$

To clarify the notation and to describe some receiver structures, we discuss a specific example. Consider a two-user scenario with symbol durations $T_1 T$ and $T_2 = 4T$, with a common chipping rate. The resulting spreading code matrix is given by

$$\tilde{\mathbf{S}} = \begin{bmatrix} \mathbf{s}_1^{(1)} & \mathbf{s}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{s}_1^{(2)} & \mathbf{0} & \mathbf{s}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{s}_1^{(3)} & \mathbf{0} & \mathbf{0} & \mathbf{s}_2 & \mathbf{0} \\ \mathbf{s}_1^{(4)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{s}_2 \end{bmatrix}$$

where

$$\mathbf{s}_1 = \begin{bmatrix} \mathbf{s}_1^{(1)} \\ \mathbf{s}_1^{(2)} \\ \mathbf{s}_1^{(3)} \\ \mathbf{s}_1^{(4)} \end{bmatrix}.$$

In this example, the second, high-rate user can be treated as four low-rate users with spreading codes:

$$\mathbf{s}_2^{(1)} = \begin{bmatrix} \mathbf{s}_2 \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{s}_2^{(2)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_2 \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \mathbf{s}_2^{(3)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{s}_2 \\ \mathbf{0} \end{bmatrix} \quad \mathbf{s}_2^{(4)} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{s}_2 \end{bmatrix}.$$

For our example, the data vector and the amplitude matrix are given by

$$\tilde{\mathbf{A}} = \mathrm{diag}\left[A_1, A_2, A_2, A_2, A_2\right]$$

$$\tilde{\mathbf{b}}(i) = \left[b_1(i), b_2(4i), b_2(4i+1), b_2(4i+2), b_2(4i+3)\right].$$

With such a view, the previously described detectors for single-rate communications can be constructed for multiple-data rate DS-CDMA systems employing variable spreading gain or variable chipping rate access. From sufficiency arguments, the observation used for detection should correspond to durations of T^*; however, we note that this approach incurs a "detection delay" for high data-rate users. Suboptimal methods can be constructed which employ detection strategies operating at the higher data rate. For our previous example, a receiver for the $4i + 2'$th bit of user 2 can be designed using the following two-user system model:

$$\bar{\mathbf{r}}(i) = \tilde{\mathbf{S}}(2)\tilde{\mathbf{A}}(2)\tilde{\mathbf{b}}^{(2)}(i) + \tilde{\mathbf{n}}^{(2)}(i) \tag{91}$$

$$\mathbf{S}(2) = \left[\mathbf{s}_1^{(3)}, \mathbf{s}_2\right] \quad \tilde{\mathbf{A}}(2) = \mathrm{diag}[A_1, A_2]$$
$$\tilde{\mathbf{b}}^{(2)}(i) = \left[b_1(i), b_2(4i+2)\right]. \tag{92}$$

Each of the receivers examined in Section V can be constructed for the signal descriptions in Eqs. (87) (low-rate signal model) and (91) (high-rate signal model). It should be emphasized that the model in Eqs. (91) should only be used for the detection of high rate users, unless used in conjunction with some kind of combining (Chen and Mitra, 1999; Saquib *et al.*, 1998). This is due to the fact that there is reduced signal energy due to the low-rate users for this model. Using these two perspectives, the following receiver structures for multirate systems have been developed: matched filter (McTiffin *et al.*, 1994), rate specific decorrelating detectors (Chen and Mitra, 1999a,b; Saquib *et al.*, 1998), minimum-mean squared error receivers (Ge and Ma, 1998), maximum likelihood receivers

(Chen and Mitra, 1999b; Mitra, 1999), and interference cancellation schemes (Johansson and Svensson, 1999).

To reveal some interesting characteristics of multirate DS-CDMA signals, we focus on the development of minimum-mean squared-error receivers. The receiver developed herein can also be applied to multicarrier CDMA systems (Lee and Milstein, 1999; Milstein, 2000). It can be shown that the optimal MMSE detector is time-varying, but periodic with period T^* (Buzzi *et al.*, 2001; Sabharwal *et al.*, 2000a,b). For the design of MMSE receivers, the desired time-varying filter will be the solution to

$$\mathbf{c}_k(i)_{\text{MMSE}} = \arg\min \mathbf{E}\left\{\left|b_k(i) - \mathbf{c}(i)^H \mathbf{r}(i)\right|^2\right\}.$$

Due its periodic nature, the optimal receiver $\mathbf{c}_k(i)_{\text{MMSE}}$, will have the following Fourier Series representation,

$$\mathbf{c}_k(i)_{\text{MMSE}} = \sum_{q=1}^{R} \mathbf{c}_k^{(q)} \exp\left(\frac{j2\pi qi}{R}\right).$$

Let R be such that $T = R \min_i T_i$. The subfilters $\mathbf{c}_k^{(q)}$ are not time-varying. Thus, we can rewrite our desired criterion in (93) as

$$\mathbf{c}_k(i)_{\text{MMSE}} = \arg\min \mathbf{E}\left\{\left|b_k(i) - \tilde{\mathbf{c}}^H \tilde{\mathbf{r}}(i)\right|^2\right\},$$

where

$$\tilde{\mathbf{c}} = \left[\mathbf{c}_k^{(0)^H}, \mathbf{c}_k^{(1)^H}, \ldots, \mathbf{c}_k^{(R-1)^H}\right]^H$$

$$\tilde{\mathbf{r}}(i) = \mathbf{r}(i) \odot \theta(i)$$

$$\theta(i) = \left[1, e^{j2\pi i}, \ldots, e^{j2\pi(R-1)i}\right]^T.$$

The Schur product operator is denoted by $\odot$; in the sequel, the Kronecker product operator will also be required, $\otimes$. The desired static receiver is given by

$$\tilde{\mathbf{c}} = \mathbf{E}[\tilde{\mathbf{r}}(i)\tilde{\mathbf{r}}(i)^H]^{-1}\mathbf{E}[b_k(i)\tilde{\mathbf{r}}(i)],$$

where

$$\mathbf{E}[b_k(i)\tilde{\mathbf{r}}(i)] = \mathbf{s}_1 \otimes \theta(i).$$

The receiver can implemented as depicted in Fig. 7. These receiver structures have been investigated for both multimedia applications and for narrowband interference suppression (Buzzi *et al.*, 2001; Sabharwal *et al.*, 2000a,b). The extension to the consideration of multipath channels is straightforward if the channel is known and the strategies for unknown channels as discussed in the prequel can be applied. The adaptive implementation of this multirate receiver is discussed in Buzzi *et al.* (2001).

D. Coded DS-CDMA

All of the receivers designed (previously cited) have not taken coding into consideration. Given the deleterious effects of the wireless channel, any practical DS-CDMA system will have some form of channel coding present. A traditional strategy for coded DS-CDMA systems was to employ a conventional receiver and then construct a Viterbi decoder with the assumption that a single-user signal is input into the decoder. Such a system would naturally suffer from multiple-access interference. An improved system could employ a multiuser detection strategy and then input the single-user coded receiver output into a decoder. However, performance will still be degraded due to the fact that the structure of the code is not taken into consideration. We shall describe two topics in the area of coded multiuser systems. We begin with the determination of an optimal sequence decoding strategy for convolutional codes and multiuser systems and then investigate optimal code design for multiuser transmit diversity systems.

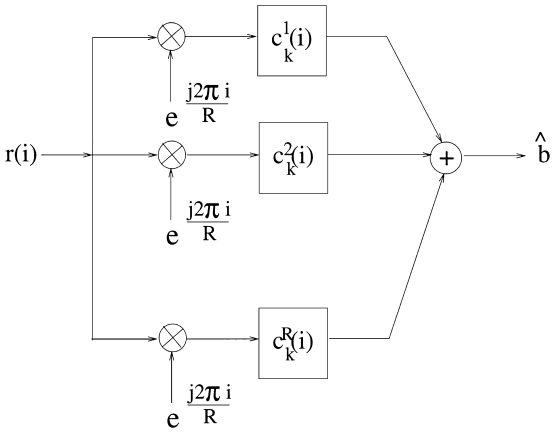


FIGURE 7 Implementation of MMSE receiver for multirate signaling.

1. Optimal Multiuser Decoders

An optimal multiuser decoder is thus described. The focus will be on convolutional codes as they operate in a sequential fashion.

A convolutional encoder can be constructed with shift-registers. The shift-register maps k_c input bits into N_c output bits resulting in a rate $R_c = k_c/N_c$ encoder. The shift-register consists of L stages with N_c modulo-2 adders. To simply describe the development of the jointly optimal multiuser decoder we consider the $R_c = \frac{1}{2}$ case. In such a scenario, the code-bit duration T is related to the information bit duration T_b by $T_b = 2T$. Note that the symbol duration is equivalent to the code-bit duration. The bit input to the encoder, $b_k(i)$ results in the following output bits $[d_k^{(1)}(i), d_k^{(2)}(i)]$. These two coded bits are then spread, modulated onto a carrier and transmitted. The contribution of the coded bits can be described using an augmented data bit description. First, some notation is necessary: Let

$$g_1(t) = \begin{cases} 1 & t \in [0, T) \\ 0 & \text{otherwise} \end{cases}$$

$$g_2(t) = \begin{cases} 1 & t \in [T, 2T) \\ 0 & \text{otherwise} \end{cases}.$$

Then let

$$\tilde{d}_k(t - 2iT - \tilde{\tau}_k) = d_k^{(1)} g_1(t - 2iT - \tilde{\tau}_k) + d_k^{(2)} g_2(t - 2iT - \tilde{\tau}_k),$$

where

$$\tilde{\tau}_k = T + \tau_k.$$

Furthermore, we define an augmented spreading waveform,

$$\tilde{s}_k(t - 2iT - \tilde{\tau}_k) = s_k(t - 2iT - \tilde{\tau}_k) + s_k\left(t - 2\left(i + \frac{1}{2}\right)T - \tilde{\tau}_k\right).$$

Now the continuous time received signal can be expressed in terms of these augmented data bits and augmented spreading waveforms,

$$r(t) = \sum_{i=-M}^{M} \sum_{k=1}^{K} A_k \tilde{d}_k(t - 2iT - \tilde{\tau}_k)\tilde{s}_k \times (t - 2iT - \tilde{\tau}_k) + n(t).$$

The optimal decoder is formed as in Section VII.A.2. The key difference is that now a four-ary symbol is to be detected and this symbol has dependence on prior symbols due to the convolutional code. Given the constraint length L, it can be shown that the four-ary super symbol $[d_k^{(1)}(i), d_k^{(2)}(i)]$ is dependent on L information symbols. As a result, the jointly optimal maximum likelihood decoder associated with the received signal can be described by a trellis with 2^{LK-1} states in contrast to 2^{K-1} states for the uncoded case.

Even with the Viterbi decoder implementation, the jointly optimal multiuser decoder incurs high computational complexity. This fact has motivated the development of several suboptimal multiuser decoders based on iterative designs. In Moher (1998), iterative techniques for cross-entropy minimization are employed to optimize the desired metric. An iterative decoder with a *maximum a posteriori* metric and a bank of K single user decoders is employed in Reed *et al.* (1998). In a related approach (Wang and Poor, 1999), a soft-input soft-output multiuser detector is used in conjunction with single-user soft-input soft-output channel decoders in an iterative fashion. Reed *et al.* (1998) and Wang and Poor (1999) use single user decoders based on the BCJR algorithm of Bahl *et al.* (1974). These iterative receivers do provide strong performance with more moderate complexity.

2. Optimal Space-Time Block Codes for DS-CDMA

We briefly treat the topic of space-time block coding for DS-CDMA systems. Methods for introducing diversity have long been used to combat fading in wireless channels (Proakis, 1995). Recently, transmit diversity techniques have received significant attention due to their efficiency in bandwidth and power (Alamouti, 1998; Guey *et al.*, 1996; Tarokh *et al.*, 1998). This early work has focused on narrowband communication systems. The schemes in Alamouti, Guey *et al.*, and Tarokh *et al.* are in part motivated by information theoretic analyses (Foschini and Gans, 1998; Telatar, 1995), which show that capacity increases linearly in the minimum of the number of transmit and receive antennae. These capacity results were shown under the assumptions of statistically independent flat Rayleigh channels between all transmit/receive antenna pairs.

The uplink between K active users and one base station is considered. The transmitter is equipped with L_t transmit antennas and the receiver is equipped with 1 receive antenna. Figure 8 depicts the encoder block diagram. The transmitter maps k_c information bits of user k, $[b_k(1), b_k(2), \ldots, b_k(k_c)]$, to one of $K_c = 2^{k_c}$ space-time block codes, $\mathcal{D}_0^{(k)} \in \mathbb{C}^{N_c \times L_t}$,

$$\mathcal{D}_0^{(k)} = \begin{bmatrix} d_1^{(k)}(1) & d_2^{(k)}(1) & \ldots & d_{L_t}^{(k)}(1) \\ d_1^{(k)}(2) & d_2^{(k)}(2) & \ldots & d_{L_t}^{(k)}(2) \\ \vdots & \vdots & \ddots & \vdots \\ d_1^{(k)}(N_c) & d_2^{(k)}(N_c) & \ldots & d_{L_t}^{(k)}(N_c) \end{bmatrix}. \quad (93)$$

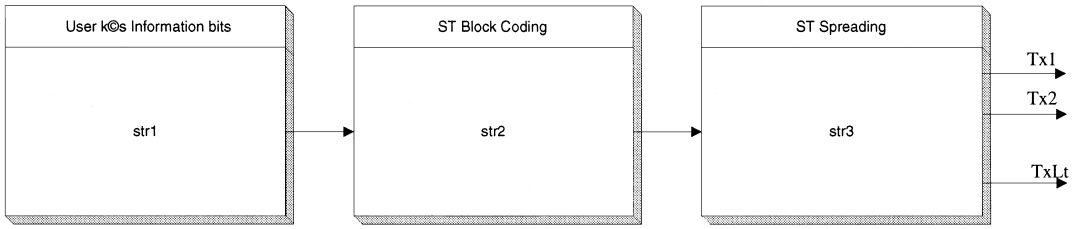


FIGURE 8 Transmitter encoder block diagram for user k.

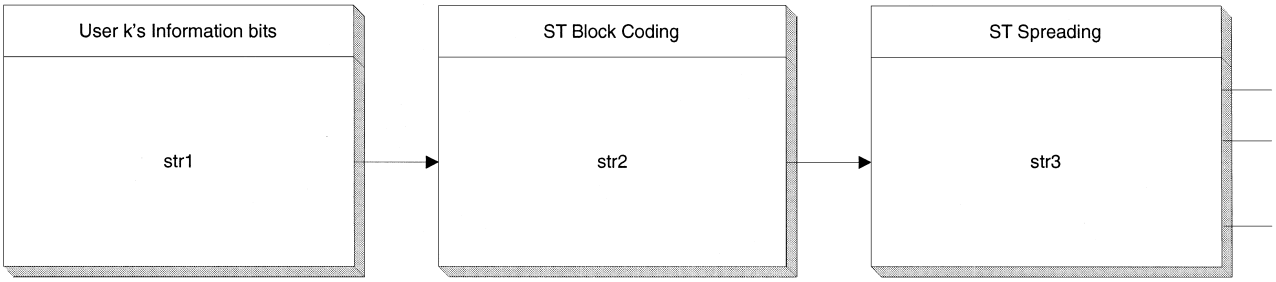


FIGURE 9 Receiver decoder block diagram.

The block length, in terms of symbol duration, is N_c. The code rate is

$$R_c = \frac{k_c}{N_c}. \tag{94}$$

Each $d_i^{(k)}(t)$ is spread by a corresponding spreading code $\mathbf{s}_i^{(k)}(t)$ and transmitted via the corresponding transmit antenna TX_i. Note that each row of the codeword matrix is transmitted simultaneously. The receiver block diagram is shown in Fig. 9. A matched filter bank, which is matched to each multipath of each user (there are L_c multipath for each user's channel), is constructed and its output forms a sufficient statistic for the decoder. The decoder provides an estimate of the information sequence for each user $k[\hat{b}_k(1), \hat{b}_k(2), \ldots, \hat{b}_l(k_c)]$.

The matched filter outputs, which form a sufficient statistic for the decoder, can be written as

$$\mathbf{y} = \mathbf{R}D\mathbf{h} + \mathbf{m} \in \mathbb{C}^{N_c L_c K L_t \times 1}, \tag{95}$$

where $\mathbf{R}$ is the spreading code correlation matrix, $\mathbf{m} \triangleq \mathbf{S}^T\mathbf{n}$ is distributed as $\mathcal{CN}(0, \sigma^2\mathbf{R})$. The diagonal matrix D, contains the coded bits for all users for the block length N_c. For the quasi-static channel,

$$\mathbf{h}(1) = \mathbf{h}(2) = \cdots = \mathbf{h}(N_c), \tag{96}$$

thus we can describe D and $\mathbf{h}$ as

$$D = \begin{bmatrix} D(1) \\ \vdots \\ D(N_c) \end{bmatrix} \in \mathbb{C}^{N_c L_c K L_t \times L_c K L_t},$$
$$\mathbf{h} = \mathbf{h}(1) \in \mathbb{C}^{L_c K L_t \times 1}. \tag{97}$$

Conditioned on the channel coefficients, the matched filter output vector is Gaussian distributed, $\mathbf{y} \sim \mathcal{CN}(\mathbf{R}D\mathbf{h}, \sigma^2\mathbf{R})$. We shall consider selecting the space-time code by optimizing the performance of the maximum likelihood decoder. The average probability of decoding D_β while D_α is transmitted is upper bounded by the Chernoff bound. In turn, for high SNR, the Chernoff bound can be approximated (Tarokh *et al.*, 1998) by

$$E_{\mathbf{h}}[P(D_\alpha \to D_\beta \mid \mathbf{h})] \le \left(\frac{1}{2\sigma}\right)^{-r} \left(\prod_{i=1}^{r} \lambda_i\right)^{-1}, \tag{98}$$

where λ_i's are the nonzero eigenvalues of $\Sigma_{\mathbf{h}}\Phi$ and r is the number of λ_i's. The matrix Φ is given by,

$$\Phi = (D_\alpha - D_\beta)^H R(D_\alpha - D_\beta)$$
$$= \Delta D^H R \Delta D \in \mathbb{C}^{N_c L_c K L_t \times N_c L_c K L_t}. \tag{99}$$

Taken over all codeword pairs, the minimum possible number of nonzero eigenvalues of Φ determines the diversity gain,

$$\Delta_H = \min_{\Delta D} \text{rank}(\Phi). \tag{100}$$

For those pairs of codewords which achieve Δ_H, the smallest product of nonzero eigenvalues of Φ usually dominates the coding gain,

$$\Delta_p = \min_{\Delta D} \prod_{i=1}^{\Delta_H} \lambda_i. \tag{101}$$

In the case of full diversity, $\Delta_p = \min_{\Delta D} \det(\Phi)$. While diversity gain and coding gain are the classical space-time code design criteria, the distance spectrum of the code set is a more accurate performance indicator (Aktas and Fitz, 2000).

A number of key observations are made after determining optimal designs for multiuser space-time coded systems employing DS-CDMA signaling:

- It can be shown that for the uplink, the space-time multiuser coding problem decouples into multiple single user space-time coding problems.
- Due to the correlation of the spreading codes used at the transmit antenna array, the optimal space-time codes for the DS-CDMA case are different from that for the narrowband case.
- Spreading improves performance.
- The optimal DS-CDMA space-time codes offer improved performance over the narrowband designs used in a DS-CDMA environment.

- For the downlink, a multiuser space-time coding problem exists and optimal designs must be done for all users simultaneously.

Other work on transmit diversity systems for DS-CDMA (Papadias, 1999; Tehrani *et al.*, 1999) employs codes constructed using narrowband design criteria or considers smart signal processing rather than coding. Space-time multiuser detection was discussed in Wang and Poor (1999) which provided iterative and adaptive decoders to exploit channel diversity.

IX. CHANNEL ESTIMATION

A number of the receivers described in Section VII.A require an estimate of the channel parameters in order to be constructed. As a result, channel estimation is an important problem to be addressed in order to build multiuser receivers. Several key channel estimation strategies are outlined in the following, with a slight emphasis on schemes based on maximum likelihood principles.

Single-path channel parameter estimation methods based on maximum likelihood (ML) schemes have been proposed in Bensley and Aazhang (1998), Smith and Miller (1994), Xie *et al.* (1993), and Zheng *et al.* (1997). In both Bensley and Aazhang (1998) and Zheng *et al.* (1997), the multiple access interference is modeled as a colored Gaussian process and ML-based estimation algorithms are developed. In Bensley and Aazhang (1998), a structured estimate is proposed for the covariance matrix of this colored Gaussian noise based on an eigenanalysis of the sample covariance matrix. In Zheng *et al.* (1997) a large sample ML estimator is presented using an approximation to the likelihood function which is asymptotically valid for a large number of bits. However, it should be noted that for most outdoor wireless applications, a frequency-selective fading model is more accurate for the channel than the flat-fading model (Proakis, 1989) investigated in the previously cited single-path schemes. In Sengupta *et al.* (1998), the large sample ML estimation method of Zheng *et al.* (1997) has been extended to the case of multiple users transmitting known training signals over multipath channels. The proposed two-stage algorithm first estimates the channel filters of all the users by employing a large sample ML technique, then extracts the channel parameters of all the paths of each user by a least-squares fit to the estimated channel filter. Approximate ML estimators based on the expectation-maximization (EM) (Dempster *et al.*, 1977) algorithm are considered in Radovic and Aazhang (1993) for joint data detection and delay estimation in asynchronous CDMA systems. A single-user, fast-fading, DS-CDMA system is considered in Sharfer and Hero (1995). However both (Radovic and Aazhang, 1993; Sharfer and Hero, 1995) suffer from being single-path techniques.

Of late, estimation algorithms which exploit signal and noise subspaces have received much attention. These subspaces are estimated using second order statistics of the received signal. Some of the earliest work for DS-CDMA systems is seen in Bensley and Aazhang (1996), where channels with multipath characteristics are considered. Simultaneously, a similar single-path blind subspace estimation technique was introduced in Ström *et al.* (1996). Both of these works estimate time delays as continuous valued unknowns. Other subpsace algorithms (Liu and Xu, 1996; Mitra *et al.*, 1997; Torlak and Xu, 1997), require coarse timing information about the user of interest and presume a fixed-spacing tap delay line model which could result in over-parameterization of the estimation problem. The multipath channel techniques proposed in Ertin *et al.* (2001) are also based on Gaussian approximations of the MAI and are iterative implementations of the maximum likelihood estimators of A key contribution is the fact that the algorithms can estimate the parameters of multiple new users.

X. POWER CONTROL

It has been noted that the conventional receiver suffers from the near–far problem. Thus, if a powerful interferer exists, a weaker user employing the conventional receiver will suffer in terms of performance. The near–far effect can be mitigated by actively controlling the transmitted power of the active users in the system. Due to the effects of fading, it may be desirable to employ power control even if the more sophisticated receivers described herein are used. Due to the complexity and variety of power control algorithms, we briefly discuss the form of such algorithms and then provide a survey of some relevant research in the area.

Let $P_k(i)$ be the transmitted power for user k. Many power control algorithms iterate until a fixed point is achieved in the power allocation scheme; the index i refers to the iteration number. A large number of decentralized power control algorithms have the following form:

$$P_k(i+1) = f(P_k(i), \mathrm{SIR}_k(i), \gamma_k),$$

where $\mathrm{SIR}_k(i)$ is the signal-to-interference ratio experienced by user k at iteration i given the power assignments of $P_1(i), \ldots P_K(i)$ for each active user in the system. The parameter γ_k is the target SIR for user k. The operator $f(\cdot)$ is the power update function which varies from algorithm to algorithm.

It has been found that the optimum power assignment is obtained by equalizing the received SINR of all mobiles, and letting each mobile's transmitted power approach infinity with these ratios. Early work (Grandhi *et al.*, 1993;

Zander, 1992) on orthogonal systems[9] established that the optimum SINR is the largest real eigenvalue of a relative gain matrix, and the optimum power level is given by the corresponding eigenvector. Similar results (Yates, 1995) were later demonstrated for DS-CDMA networks.

The power control problem often reduces to finding the maximum set of mobiles that can be served. Early research (Zander, 1992) suggested centralized iterative methods to drop existing mobiles and reestablish the target SINR. Distributed methods followed in Grandhi *et al.*, 1994; Lee *et al.* (1995), and Zander (1992). The effect of receiver noise was considered in Foschini and Miljanic (1993, 1995), resulting in an algorithm that incorporates power control into the local autonomous dynamic channel allocation procedure so that both power control and channel access are performed in a distributed manner with local measurements and decision making. Finally, we note that power control problems have also been studied in the context of multiuser detection (Ulukus and Yates, 1998; Yener *et al.*, 1999).

XI. CONCLUSIONS

In this chapter, multiuser systems based on Direct-Sequence Code-Division Multiple-Access are introduced. A historical perspective is provided. After the introduction of a simple signal model, several key multiuser receivers are described. Performance measures specialized to the multiuser scenario are described and determined for some of the basic multiuser receivers. The bulk of the chapter is devoted to the description of receivers specialized to several particular communication scenarios and receiver mechanisms: adaptive algorithms, reduced rank methods, multimedia receivers, and coded systems. The chapter concludes with the discussion of two problem areas integral to the successful deployment of a DS-CDMA system: channel estimation and power control.

SEE ALSO THE FOLLOWING ARTICLES

MULTIPLEXING • RADIO PROPAGATION • SIGNAL PROCESSING, DIGITAL • WIRELESS COMMUNICATIONS

BIBLIOGRAPHY

Adachi, F., Sawahashi, M., and Suda, H. (1998). "Wideband DS-CDMA for next-generation mobile communication systems," *IEEE Commun. Magazine* 56–69.

Aktas, D. K., and Fitz, M. P. (2000). "Computing the distance spectrum of space-time trellis codes." *In* "Proceedings WCNC'00," Chicago IL.

Alamouti, S. M. (1998). "A simple transmit diversity technique for wireless communications," *IEEE Trans. Select Areas Comm.* **16**(8), 1451–1458.

Bahl, L. R., Cocke, J., Jelenik, F., and Raviv, J. (1974). "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE. Trans. Inform. Theory* **20,** 284–287.

Barbosa, A. N., and Miller, S. L. (1998). "Adaptive detection of DS/CDMA signals in fading channels," *IEEE Trans. Commun.* **46**(5), 115–124.

Bensley, S. E., and Aazhang, B. (1998). "Maximum likelihood synchronization of a single user for CDMA communication systems," *IEEE Trans. Commun.* **46**(3), 392–399.

Bensley, S. E., and Aazhang, B. (1996). "Subspace-based channel estimation for code division multiple access communication systems," *IEEE Trans. Commun.* **44**(8), 1009–1020.

Buzzi, S., Lops, M., and Tulino, A. (2001). "Partially blind adaptive MMSE interference rejection in asynchronous DS/CDMA networks over frequency selective fading channels," *IEEE Trans. Commun.* In press.

Chapin, E. W. (1960). "The past and future techniques of vehicular communications," *IRE Trans. Veh. Technol.* **VC-9**(2), 6–10.

Chen, J., and Mitra, U. (1999a). "Analysis of decorrelator-based receivers for multi-rate CDMA communications," *IEEE Trans. Veh. Technol.* **6,** 1966–1983.

Chen, J., and Mitra, U. (1999b). "Dual-rate MMSE receivers for DS/CDMA signals: Analysis and comparisons," *IEEE Trans. Info. Theory* **45**(7), 2434–2447.

Chen, W., Mitra, U., and Schniter, P. (2000). "Reduced rank detection schemes for ds-cdma communication systems," *IEEE Trans. Inform. Theory*. Submitted for Publication.

Cooper, G. R., and Nettleton, R. W. (1978). "A spread-spectrum technique for high-capacity mobile communications," *IEEE Trans. Veh. Technol.* **27**(4), 264–275.

Dahlman, E., Gudmundson, B., Nilsson, M., and Sköld, J. (1998). "UMTS/IMT2000 based on wideband CDMA," *IEEE Commun. Magazine* 70–80.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B.* 1–38.

Duel-Hallen, A. (1992). "Equalizers for multiple input/multiple output channels and PAM systems with cyclostationary input sequences," *IEEE Trans. Commun.* **10**(3), 630–639.

Ström, E. G., Parkvall, S., Miller, S. L., and Ottersten, B. E. (1996). "Propagation delay estimation in asynchronous direct-sequence code-division multiple access systems," *IEEE Trans. Commun.* **44**(1), 84–93.

Ertin, E., Mitra, U., and Siwamogsatham, S. (2001). "Maximum-likelihood based multipath channel estimation for code-division multiple-access systems," *IEEE Trans. Commun.* **49**(2), 290–302.

Foschini, G. J., and Miljanic, Z. (1993). "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Veh. Technol.* **42**(4), 641–646.

Foschini, G. J., and Miljanic, Z. (1995). "Distributed autonomous wireless channel assignment algorithm with power control," *IEEE Trans. Veh. Technol.* **44**(3), 420–429.

Foschini, G. J., and Gans, M. J. (1998). "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.* **6,** 311–335.

Freret, J. P. (1980). "Applications of spread-spectrum radio to wireless terminal communications," *Proc. of NRC '80* **4,** 69.7.1–69.7.4.

Ge, H., and Ma, J. (1998). "Multi-rate LMMSE detectors for asynchronous multi-rate CDMA systems." *In* "Proceedings of the

[9]We use the term *orthogonal* to refer to TDMA- and FDMA-based systems.

1998 IEEE International Conference on Communication, ICC," volume 2, pp. 714–718, Atlanta, GA.

Gold, R. (1967). "Optimum binary sequences for spread-spectrum multiplexing," *IEEE Trans. Info. Theory* **13,** 619–621.

Goldstein, J. S., and Reed, I. S. (1997). "Reduced rank adaptive filtering," *IEEE Trans. Sig. Processing* **45**(2), 492–496.

Goldstein, J. S., Reed, I. S., and Scharf, L. L. (1998). "A multistage representation of the Wiener filter based on orthogonal projections," *IEEE Trans. Info. Theory* **44**(7), 2943–2959.

Golub, G. H., and Van Loan, C. F. (1989). "Matrix Computations," 2nd edition, Johns Hopkins University Press, Baltimore, MD.

Grandhi, S. A., Goodman, D. J., Vijayan, R., and Zander, J. (1993). "Centralized power control in cellular radio systems," *IEEE Trans. Veh. Technol.* **42**(4), 466–468.

Grandhi, S. A., Vijayan, R., and Goodman, D. J. (1994). "Distributed power control in cellular radio systems," *IEEE Trans. Commun.* **42**(2–4), 226–228.

Guey, J. C., Fitz, M. P., Bell, M. R., and Kuo, W. Y. (1996). "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," *Proc. IEEE Veh. Technol. Conf.* **1,** 136–140.

Haardt, M., Klein, A., Koehn, R., Oestreich, S., Purat, M., Sommer, V., and Ulrich, T. (2000). "The TD-CDMA based UTRA TDD mode," *IEEE J. Select. Areas Commun.* **18**(8), 1375–1386.

Hartley, R. (1928). "Transmission of information," *Bell. Sys. Tech. J.* **7**(3), 535–563.

Haykin, S. (1996). "Adaptive Filter Theory," Prentice-Hall, NJ.

Honig, M. L., and Goldstein, J. S. (2000). "Adaptive reduced-rank interference suppression based on the multi-stage Weiner filter," in preprint.

Honig, M. (1995). "Upamanyu Madhow, and Sergio Verdú. Blind adaptive multiuser detection," *IEEE Trans. Inform. Theory* **41**(4), 944–960.

Horn, R. A., and Johnson, C. R. (1985). "Matrix Analysis," Cambridge University Press, Cambridge.

Huang, H., and Papadias, C. (1998). "Linear space-time multiuser receivers for wireless cdma systems." *In* "Ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications," volume 1, pages 238–242, September 1998.

Johansson, A., and Svensson, A. (1999). "On multirate DS-CDMA systems with interference cancellation," *Wireless Pers. Commun.* **9**(1), 1–29.

Kasami, T., Lin, S., and Peterson, W. (1968). "Some results on cyclic codes which are invariant under the affine group and their applications," *Inf. Control* **11,** 475–496.

Lee, D., and Milstein, L. B. (1999). "Comparison of multicarrier ds-cdma broadcast systems in a multipath fading channel," *IEEE Trans. Commun. Year*.

Lee, T.-H., Lin, J.-C., and Su, Y. T. (1995). "Downlink power control algorithms for cellular radio systems," *IEEE Trans. Veh. Technol.* **44**(1), 89–94.

Liu, H., and Xu, G. (1995). "Closed-form blind symbol estimation in digital communications," *IEEE Trans. Signal Process.* **43**(11), 2714–2723.

Liu, H., and Xu, G. (1995). "A subspace method for signature waveform estimation in synchronous CDMA systems," *IEEE Trans. Commun.* 1346–1354.

Liu, H., and Zoltowski, M. (1997). "Blind equalization in antenna array CDMA systems," *IEEE Trans. Signal Process.* **45**(1), 161–172.

Lupas, R., and Verdú, S. (1989). "Linear multiuser detectors for synchronous Code-Division Multiple-Access channels," *IEEE Trans. Inform. Theory* **35**(1), 123–136.

Lupas, R., and Verdú, S. (1990). "Near-far resistance of multi-user detectors in asynchronous channels," *IEEE Trans. Commun.* **38,** 496–508.

Madhow, U., and Honig, M. L. (1994). "MMSE interference suppression for direct-sequence spread spectrum CDMA," *IEEE Trans. Commun.* **42**(12), 3178–3188.

Magnuski, H. (1961). "Wideband channel for emergency communication," *IRE Int. Conv. Rec.* 80–84, Mar. 1961, note Mar. 20–23, part 8.

McTiffin, M. J., Hulbert, A. P., Ketseoglou, T. J., Heimsch, W., and Crisp, G. (1994). "Mobile access to an ATM network using a CDMA air interface," *IEEE J. Select. Areas Commun.* **12**(5), 900–908.

Miller, M. L., Honig, S. L., and Milstein, L. B. (2000). "Performance analysis of MMSE receivers for DS-CDMA in frequency-selective fading channels," in preprint.

Milstein, L. B. (2000). "Wideband code division multiple access," *IEEE J. Select. Areas Commun.* **18**(8), 1344–1354.

Mitra, U. (1999). "Comparison of maximum-likelihood based detection for two multi-rate access schemes for CDMA signals," *IEEE Trans. Commun.* **47**(1), 64–77.

Mitra, U., Slock, D., and Escudero, C. (1997). "Blind equalization techniques for CDMA communications." *In* "Proceedings of the 31st Annual Conference on Information Sciences and Systems," Johns Hopkins, Bethesda, MD.

Moher, M. (1998). "An iterative multiuser decoder for near-capacity communications," *IEEE Trans. Commun.* **46,** 870–880.

Moshavi, S., Kanterakis, E. G., and Schilling, D. L. (1996). "Multistage linear receivers for DS-CDMA systems," *Int. J. Wireless Inf. Networks* **3**(1), 1–17.

Nyquist, H. (1924). "Certain factors affecting telegraph speed," *Bell. Sys. Tech. J.* **3**(2), 324–346.

Ojanperä, T., and Prasad, R. (1998). "An overview of air interface multiple access for IMT-2000/UMTS," *IEEE Commun. Magazine* 82–95.

Okamura, Y., and Adachi, F. (1998). "Variable-rate data transmission with blind rate detection for coherent ds-cdma mobile radio," *IEICE Trans. Commun.* E81-B:71365–1373.

Pados, D. A., Lombardo, F. J., and Batalama, S. N. (1999). "Auxiliary-vector filters and adaptive steering for DS/CDMA single-user detection," *IEEE Trans. Veh. Technol.* **48**(6), 1831–1839.

Papadias, C. (1999). "On the spectral efficiency of space-time spreading schemes for multiple antenna CDMA systems." *In* "Proc. 33rd Asilomar Conference on Signals, Systems and Computers," Pacific Grove, CA.

Papoulis, A. (1984). "Probability, Random Variables, and Stochastic Processes," 2nd edition, McGraw Hill, New York.

Patel, P., and Holtzman, J. (1994). "Analysis of a simple successive interference cancellation scheme in a DS/CDMA system," *IEEE J. Select. Areas Commun.* **12**(5), 796–807. Issue on Code-Division Multiple-Access Networks, II.

Pickholtz, R. L., Schilling, D. L., and Milstein, L. B. (1982). "Theory of spread spectrum communications—A tutorial," *IEEE Trans. Commun.* **30**(5), 855–884.

Pierce, J. R. (1949). "Time division multiplex system with erratic sampling times," Technical report, Bell Telephone Laboratories.

Poor, H. V. (1994). "An Introduction to Signal Detection and Estimation," 2nd edition, Springer-Verlag, New York.

Porat, B. (1994). "Digital Processing of Random Signals: Theory and Methods," Information and System Sciences Series, Prentice-Hall, Englewood Cliffs, NJ.

Proakis, J. G. (1995). "Digital Communications," 3rd edition, McGraw-Hill, New York.

Proakis, J. G. (1989a). "Digital Communications," 2nd edition, McGraw Hill Series in Communications and Signal Processing, McGraw Hill, New York.

Proakis, J. G. (1989b). "Digital Communications," 2nd edition, McGraw Hill Series in Communications and Signal Processing, McGraw Hill, New York.

Proakis, J. G., and Salehi, M. (1994). "Communication Systems Engineering," Prentice-Hall, Englewood Cliff, NJ.

Radovic, A., and Aazhang, B. (1993). "Iterative algorithms for joint data detection and delay estimation for CDMA communication systems." *In* "Proc. of the 31st Annual Allerton Conference on Communication, Control, and Computing," Allerton House, Monticello, IL.

Rappaport, T. (1996). "Wireless Communications: Principles & Practice," Prentice-Hall, Englewood Cliffs, NJ.

Reed, M. C., Schlengel, C. B., Alexander, P. D., and Asenstorfer, J. A. (1998). "Iterative multiuser detection for cdma with fec: Near-single-user performance," *IEEE Trans. Commun.* **46,** 1693–1699.

De Rosa, L. A., and Rogoff, M. (1950). "Sect. 1 (Communications) of "Application of Statistical Methods to Secrecy Communcations Systems," Proposal 946, Fed. Telecommun. Lab., Nutley, NJ.

Sabharwal, A., Mitra, U., and Moses, R. (2000a). "Low complexity mmse receivers for multirate DS-CDMA systems." *In* "Proc. of the 2000 Conference on Information Sciences and Systems," volume 1, pp. TA3–18, Princeton, NJ.

Sabharwal, A., Mitra, U., and Moses, R. (2000b). "Mmse receivers for multirate DS-CDMA systems," *IEEE Trans. Commun.* Submitted for publication.

Saquib, M., Yates, R., and Mandayam, N. (1998). "Decorrelating detectors for a dual rate synchronous DS/CDMA channel," *Wireless Pers. Commun.* **9**(3), 197–216.

Sarwate, D. V., and Pursley, M. B. (1980). "Crosscorrelation properties of pseudorandom and related sequences," *Proc. IEEE* **68**(5), 593–619.

Schneider, K. S. (1979). "Optimum detection of Code Division multiplexed signals," *IEEE Trans. Aerospace Electr. Syst.* **16,** 181–185.

Scholtz, R. A. (1982). "The origins of spread-spectrum communications," *IEEE Trans. Commun.* COM-**30,** 822–854.

Scholtz, R. A. (1994). "The evolution of spread-spectrum multiple-access communications." *In* "Proceedings of IEEE ISSSTA'94," volume 1, pp. 4–13.

Sengupta, C., Hottinen, A., Cavallaro, J. R., and Aazhang, B. (1998). "Maximum likelihood multipath channel parameter estimation in CDMA systems," Rice University Technical Report.

Sharfer, I., and Hero, A. O. (1995). "Spread spectrum sequence estimation and bit synchronization using an EM-type algorithm." *In* "Proc. of ICASSP," volume 3, pp. 1864–1867.

Simon, K., Omura, J., Scholtz, R., and Levitt, B. (1985). "Spread Spectrum Communications," volume III, 2nd edition, Computer Science Press, New York.

Singh, R., and Milstein, L. B. (1999). "Interference suppression for DS/CDMA," *IEEE Trans. Commun.* **47**(3), 446–453.

Smith, R. F., and Miller, S. L. (1994). "Code timing estimation in a near-far environment for direct-sequence code-division multiple access." *In* "Proc. of IEEE Military Communications Conference," volume 1, pp. 47–51.

Sorelius, J., Gesbert, D., Stoica, P., and Paulraj, A. (1998). "Multi-user MMSE detection of CDMA signals: Blind and semi-blind algorithms." *In* "Proceedings of the 8th IEEE Digital Signal Processing Workshop," Bryce Canyon, UT.

Stark, H., and Woods, J. W. (1994). "Probability, Random Processes, and Estimation Theory for Engineers," 2nd edtion, Prentice Hall, Englewood Cliffs, NJ.

Stüber, G. L. (1996). "Principles of Mobile Communication," Kluwer Academic Publishers, Boston, MA.

Tarokh, V., Seshadri, N., and Calderbank, A. R. (1998). "Space-time codes for high data rate wireless communication: Performance criterion and code construction." *IEEE Trans. Inf. Theory*. **44**(2),744–765.

Tehrani, A. M., Negi, R., and Cioffi, J. (1999). "Space-time coding over a code division multiple access system." *In* "IEEE WCNC'99," New Orleans, LA.

Telatar, I. E. (1995). "Capacity of multi-antenna gaussian channels," *Technical Memorandum,* Bell Laboratories, Lucent Technologies.

Torlak, M., and Xu, G. (1997). "Blind multiuser channel estimation in asynchronous CDMA systems," *IEEE Trans. Signal Process.* **45**(1), 137–147.

Tsatsanis, M. K., and Giannakis, G. B. (1997). "Blind estimation of direct sequence spread spectrum signals in multipath," *IEEE Trans. Signal Process.* **45**(5), 1241–1252.

Ulukus, S., and Yates, R. D. (1998). "Adaptive power control and mmse interference suppression," *ACM Wireless Networks* **4**(6), 489–496.

Varanasi, M., and Aazhang, B. (1990). "Multistage detection in asynchronous CDMA communication," *IEEE Trans. Commun.* **38**(4), 509–519.

Verdú, S. (1986). "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. Inform. Theory* **32**(1), 85–96.

Verdú, S. (1986). "Optimum multi-user asymptotic efficiency," *IEEE Trans. Commun.* **34**(9), 890–897.

Verdú, S. (1998). "Multiuser Detection," Cambridge University Press, Cambridge, UK.

Viterbi, A. J. (1995). "CDMA: Priciples of Spread Spectrum Communication," Wireless Communications Series, Addison-Wesley, Reading, MA.

Wang, X., and Poor, H. V. (1998a). "Blind multi-user detection: A subspace approach," *IEEE Trans. Info. Theory* **44,** 677–690.

Wang, X., and Poor, H. V. (1998b). "Blind equalization and multiuser detection in dispersive CDMA channels," *IEEE Trans. Commun.* **46,** 91–103.

Wang, X., and Poor, H. V. (1999a). "Iterative (turbo) soft interference cancellation and decoding for coded cdma," *IEEE Trans. Commun.* **47**(7), 1046–1061.

Wang, X., and Poor, H. V. (1999b). "Space-time multiuser detection in multipath cdma channels," *IEEE Trans. Signal Process.* **47**(9), 2356–2374.

Woodward, G., and Vucetic, B. S. (1998). "Adaptive detection for DS-CDMA," *Proceedings IEEE* **86**(7), 1413–1434.

Wu, T.-H., and Geraniotis, E. (1994). "CDMA with multiple chip rates for multi-media communications." *In* "Proceedings of the Twenty-eighth Annual Conference on Information Sciences and Systems," Princeton University, NJ, pp. 992–997.

Xie, Z., Rushforth, C. K., Short, R., and Moon, T. K. (1993). "Joint signal detection and parameter estimation in multiuser communications," *IEEE Trans. Commun.* **41,** 1208–1216.

Xie, Z., Short, R. T., and Rushforth, C. K. (1990). "A family of suboptimum detectors for coherent multiuser communications," *IEEE J. Select. Areas Commun.* **8**(4), 683–690.

Yates, R. D. (1995). "A framework for uplink power control in cellular radio systems," *IEEE J. Select. Areas Commun.* **13**(7), 1341–1347.

Yener, A., Yates, R. D., and Ulukus, S. (1999). "Joint power control, multiuser detection and beamforming for CDMA systems." *In* "Proc. Vehicular Technology Conference."

Zander, J. (1992). "Distributed co-channel interference control in cellular radio systems," *IEEE Trans. Veh. Technol.* **41,** 305–311.

Zander, J. (1992). "Performance of optimum transmitter power control in cellular radio systems," *IEEE Trans. Veh. Technol.* **41**(1), 57–62.

Zheng, D., Li, J., Miller, S. L., and Strom, E. G. (1997). "An efficient code-timing estimator for DS-CDMA signals," *IEEE Trans. Signal Process.* 82–89.

Zhu, L. J., Madhow, U., and Galup, L. (2000). "Differential MMSE: two applications to interference suppression for direct-sequence CDMA," in preprint.

Ziemer, R. E., and Peterson, R. L. (1985). "Digital Communications and Spread Spectrum Systems," Houghton Mifflin, Boston, MA.



Telecommunications

John N. Daigle
University of Mississippi

Katherine H. Daigle
KHD Systems

I. Primary Networks and Technologies
II. Basic Point-to-Point Communication Systems
III. Switched Connections and Networks
IV. Early Evolution of the Telephone Network
V. Early Data Communication Networks
VI. Common Channel Signaling in the Telephone Network
VII. Recent Developments in Telecommunications
VIII. Evolution of the United States Telecommunications Industry: 1837–2000
IX. The Twenty-First Century
X. Telecommunications Timeline: Telegraph, Telephone, Television, Computer, Internet: 1837–2001

GLOSSARY

Baseband The collection of frequencies over which a signal exists in its natural setting; for example, the baseband of a voice signal is the range of frequencies below 20 KHz or 20,000 cycles/sec.

Carrier system A transmission facility that provides multiple equivalent circuits between two geographically distant points. There are two basic classes of carrier systems, short haul and long haul, designed to connect end points at less than or more than about fifty miles, respectively.

Circuit switching A method of communication in which a physical circuit is established between two end systems before communication begins to take place. In circuit switching, specific resources are set aside for the exclusive use of the end systems for the duration of the connection.

Communication circuit The set of equipment and transmission facilities that are used to provide information transmission between two end systems.

Computer communication network A collection of applications hosted on different end systems and interconnected by an infrastructure that provides intercommunications. The infrastructure consists of computer-based end systems, packet switches, and transmission facilities connecting end systems to switches and switches to each other. The Internet is a computer communication network.

End system The equipment used directly by an end user

to facilitate communications; examples include a telephone, a television, or a personal computer.

Equipment A collection of electrical components and circuitry put together in a single physical package. Examples of equipments are telephones, televisions, telephone channel banks, test sets, and radios. This descriptive word is commonly used in any systems engineering discipline in preference to the word *component* to avoid confusion. For example, systems engineers would describe a radio as equipment, but transistors, power cords, transformers, and speakers in a radio as components.

Message switching A service technology used to transfer messages between end systems on a message-by-message basis. Immediate delivery is not guaranteed, and there is no notion of a direct connection between end systems. Message switching is an electronic analog of the postal system.

Packet switching A method of communication in which messages are exchanged between end systems via the exchange of a sequence of fragments of the messages called packets. Specific resources required for transmission of packets are requested as needed during the call. Availability of the resources requested is not guaranteed.

Router This is another name for a packet or message switch. Sometimes a router is called an intermediate system.

SS7 network The packet-switched network that provides communications and control of the call-carrying, circuit-switched telephone network.

Switch A system consisting of collections of input ports, output ports, switching circuitry, and control software that provides the capability to transfer information arriving at an input port to an arbitrary output port or subset of the output ports. A circuit switch literally connects an incoming trunk to an outgoing trunk electronically for the duration of a call. A packet switch transfers information packet-by-packet to possibly different outputs based on addressing information contained in the packet.

Tariff A set of rules governing the pricing structure for a collection of services.

TCP/IP The combination of the Transmission Control Protocol and the Internet Protocol; TCP provides the rules and mechanisms for exchange of information between end systems, and the IP protocol determines how packets are routed through a communication network. TCP/IP is the protocol of choice in the Internet.

Telephone network The world-wide network that provides circuit switched connections between arbitrary end systems around the world, primarily over circuits suitable for telephone conversations and low speed data.

T1 carrier system A short-haul digital transmission system consisting of a collection of channel banks as terminal equipment and digital transmission facilities between the channel banks. Each channel bank provides input and output circuitry to support 24 voice calls, and the digital lines between the channel banks operate at 1.544 megabits/sec. A T1 system transmits between the channel banks over copper, twisted-pair cable, features baseband transmission, and has repeaters spaced at about one-mile intervals.

Trunk A circuit between two circuit switches that provides the capacity to carry the equivalent of one telephone call in one direction.

World-wide web The collection of web sites and web browsers interconnected over the world-wide Internet.

TELECOMMUNICATIONS literally means *communications at a distance*, the usual connotation being *communications at a distance by electrical means*. The array of public and private and telecommunication systems and services spanning the globe today is the result of amazing technological advances that have taken place over the last two centuries. This article focuses on those aspects of telecommunications concerned with providing the public telecommunications services that a growing proportion of the world's population uses every day.

I. PRIMARY NETWORKS AND TECHNOLOGIES

The purpose of a telecommunication system is to exchange information among users of the system. This information exchange can take place in a variety of ways, for example, multiparty voice communications, television, electronic mail, and electronic message exchange.

The world-wide telecommunications infrastructure in the year 2001 consists largely of two interrelated major infrastructures: the telephone network, which is a *circuit-switched* network, and the Internet, which is a *packet-switched computer communications* network. Today, control of the telephone network occurs mostly over computer communication networks, and transmission facilities that interconnect the major components of the Internet are derived from the same facilities that interconnect telephone switches. The state of the art today is the result of the tremendous growth in scientific knowledge and significant advances over the last two centuries of a wide variety of technologies. Indeed, it is not easy to identify areas of science and technology that do not have some impact upon

the state of the art in telecommunications. Any brief discussion of telecommunications must, then, of necessity, omit significant portions of the story.

The telecommunications infrastructure of today provides two fundamentally different types of service: message-oriented and real-time communication-oriented. Examples of message-oriented services include file transfer and inquiry-response. Such a service is naturally implemented over a computer communications network using *packet switching*. Examples of real-time communications are voice telecommunications and television, which are more naturally implemented over the telephone network using *circuit switching* (television is, of course, broadcast locally, but national programming reaches local stations over permanent circuit-switched connections). However, most services that naturally fit into one of these paradigms are provided over the other to some limited extent today.

Since at least as early as 1975, both consumers and providers of telecommunications services have been dreaming of a day when telecommunications services can be obtained in a unified way. As the twenty-first century dawns, both packet- and circuit-switched networks enjoy ubiquitous deployment, many believe this dream of integrating services may become a reality within the next few years. Among those who believe there will be exactly one telecommunications infrastructure, there are credible views based on each technology.

Telecommunications systems, be they packet- or circuit-switched, have a number of elements in common. Those elements are introduced in the following section.

II. BASIC POINT-TO-POINT COMMUNICATION SYSTEMS

Telecommunications was born out of the need that occurs in ordinary life to communicate messages over distances at a reasonable cost. This was an economic need in the most basic sense; survival depended upon the reliable transfer of the message. For example, in the case of a battlefield situation, the reliability of having messages delivered by runners crossing battle lines would be relatively low and the cost would be high; multiple runners would be required and many of them would lose their lives.

Once ways of encoding messages had been invented and a basic understanding of electrical circuits was developed, the way was paved for electrical communications. Electrical wires could be strung between two distant points. Through the process of opening and closing switches, or making and breaking contacts, hisses and pops could be sent and then interpreted as different signals. Using such a means, any desired message could be sent without actually having to carry the physical message. Furthermore, delivery could be fast and more or less private, and receipt could be verified.

To use this astounding technological advance in conveying messages, all one needed was a wire strung between the two distant points, a simple mechanism having the appropriate switches, and a trained person at each end. Samuel Morse invented such a mechanism in 1837. He later invented a scheme for encoding messages, and then, under contract with the government of the United States of America, built the first telegraph system in 1843.

The distances over which signals can be transmitted over electrical circuits without excess quality degradation are limited. More distant places could be reached by means of relaying. A mechanism for automatically relaying messages was introduced, also with Morse's involvement, soon after the introduction of the first telegraph system, thus allowing a single pair of operators to handle the entire message transfer. Acknowledgements could then be sent from the most distant receiving end to the sending end, thus verifying message delivery. Obviously, much could be gained from automating the process of repeating the messages at intermediate points.

This point-to-point system between the end points can be thought of as a basic communications channel. Among its many properties, a channel has a transmitter-receiver pair at each end and a capacity, which defines the volume of information that can be conveyed over the channel.

Once the system is in place to take care of necessary communications, any excess capacity could be used for messages having lower priority; that is, messages for which people would place a lower value of delivery. Spreading the cost among more messages would then lower the per-message cost, and this would increase the amount demanded. To further increase the amount demanded, the per-message cost would have to be lowered. The first step along the way was to raise the number of messages that could be sent simultaneously over the same wire from one to two. Next the capacity could be raised to four, and so forth. The practice of using the same transmission system for multiple simultaneous communications is called *multiplexing*, and the system that implements multiplexing is called a *multiplexing system*.

A multiplexing system supports a number of *communication channels*, each of which carries the equivalent of one conversation. A multiplexing system consists of two terminal equipments, one for each end of the connection, and an electrical transmission circuit that interconnects the terminal equipments. Each terminal equipment has a *drop* side, which faces individual *channels*, and a *line* side, which faces the transmission circuit between the two terminal equipments. In addition, each terminal equipment has a multiplexor, which combines signals of individual

channels for transmission over the line side, and a demultiplexer, which separates the signal received over the line side for transmission over the individual channels of the drop side. The input (output) circuitry associated with each channel at each multiplexer is called an *input* (*output*) *port*.

The initial multiplexing systems were based on *frequency division multiplexing*. In a frequency division multiplexing system, there is a collection of single frequency signals called *carriers*. The signal for each call to be carried, for example, the signal from a telegraph transmitter, enters the multiplexor via one of the ports on the drop-side of the multiplexor. This signal, which is called the baseband signal, modulates one of the carriers. The multiplexor then combines the collection of modulated carriers into a single electrical signal that is sent over the transmission line connecting the multiplexing equipment. At the receiving end, the demultiplexer separates the combined line signal, through the use of demodulating and filtering circuitry, into the signals that belong to each of the individual channels and then routes the signal to the correct output port.

Because frequency division multiplexing systems are based on carrier signals, they are called *carrier systems*. A carrier system is analogous to the radio broadcast system, where each radio station uses a different carrier frequency and the listener tunes to the desired station. The electrical signal representing the voice messages and music broadcast over each radio channel is the baseband signal. And, when we tune into "870 on the radio dial," we are tuning to the station whose carrier frequency is 870 KHz. The airways provide the communication link between the radio stations and the radios.

Thus, a carrier system consists of terminating equipment that is used to combine and separate signals and an electrical circuit of some kind between its end points. Individual signals enter the multiplexing systems via input ports and exit from the opposite end via output ports. Multiple parallel communication channels can be achieved over the same electrical circuit through the use of carrier systems.

Generally speaking, if a carrier system is designed to interconnect switches over a great distance, then that carrier system is called a *long-haul carrier system*, otherwise, it is called a *short-haul carrier system*. Both short- and long-haul carrier systems can have a variety of capacities in terms of the number of calls that they can handle simultaneously. In fact, the capacities of the various carrier systems range from a few circuits to many thousands of circuits.

With increased multiplexing capability, a natural improvement would be in the area of message entry and delivery. Once a machine could fill the available channels, one would want to increase the number of channels by using better circuitry at the end points or better wires and so forth. This process could continue indefinitely with advances in wire manufacturing technology as well as transmitter and receiver circuitry, including the circuitry in repeaters at intermediate points. This explains, to a large extent, the tremendous investment in basic and applied research in so many different areas of technology by the telecommunications industry.

With increased message carrying capability, the per-message cost would drop and quantity demanded would increase. This, too, would go on in a continuous cycle. In addition, it is not hard to see that significant savings in cost might be achieved by replacing wires by radio transmission and receiving equipment. For example, to send signals across a large body of water, it may be less costly to use radios than to develop and deploy underwater cables. Guglielmo Marconi developed such a radio system in the mid-1890s.

Once a carrier system is in place, it is natural to separate the location of the end-user equipment from that of the carrier system's terminating equipment. The user's equipment is then connected to the carrier system through an access line, which is often referred to as a *local loop*.

So far, then, we have a point-to-point system for transmitting messages. This system consists of the end-user equipment, a local loop at each end, and a carrier system. The carrier system, in turn, consists of its terminating equipment, repeaters needed between the terminating equipment, and electrical circuits interconnecting the repeaters and the terminating equipment.

In the case of telegraphy, the objective of the system is to convey messages from some source to some destination, much as the objective of the postal service is to deliver letters. The primary difference is that the objective is accomplished electronically, and therefore, more quickly. The *service model* is that the system takes in messages at one end and delivers them to the other end; the service itself is *message delivery*.

III. SWITCHED CONNECTIONS AND NETWORKS

Once point-to-point connections are understood, it is not difficult to see why one would want to build a network so that point-to-point connections are not required between every pair of end points. For example, to cover connections among five major cities in the United States, say, New York, St. Louis, Dallas, Seattle, and Los Angeles, there are many options. One option would be to build one carrier system between each possible pair of the five cities and deliver each message using exactly one carrier system. A second option would be to build one carrier system each

between New York and St. Louis, St. Louis and Dallas, Dallas and Seattle, Dallas and Los Angeles. In the second option, messages from New York bound for Los Angeles would go first to St. Louis, then to Dallas, then to Los Angeles; that is, the messages would be delivered by a network of carrier systems.

In the previous case, each message coming off of the carrier system would be examined to determine its appropriate next hop, and the messages could be routed hop-by-hop to their destinations. The service rendered would still be message delivery, but it would be done by a message relaying, or *message switching*, system. This system is the electronic equivalent of a postal delivery system.

A second option would be to provide a means of cross-connecting channels at the intermediate points to create a semipermanent connection between each possible source and destination pair. For example, consider the connection from New York to Los Angeles mentioned above. In St. Louis and Dallas, each input and output port of the carrier system would be terminate at a cross-connect. If there are twelve channels per carrier system, then a cross-connect would have 24 *appearances* for each carrier system, 12 for input and 12 for output. Suppose the New York to Los Angeles message stream is to use channel 1 from New York to St. Louis, channel 4 from St. Louis to Dallas, and channel 7 from Dallas to Los Angeles. Then, on the cross-connect in St. Louis, a patch cord would be placed between the channel 1 output appearance of the New York to St. Louis system and the channel 4 input of the St. Louis to Dallas system. Similarly, on the cross-connect in Dallas, a patch cord would be placed between the channel 4 output appearance of the St. Louis to Dallas system and the channel 7 input of the Dallas to Los Angeles system.

This second option provides essentially the same effect as having a point-to-point carrier system between each city pair, but its implementation better utilizes available capacity because there is broad flexibility in assigning circuits between each city pair. This option provides the same service as before (message delivery), but it provides the service over semipermanent *circuit-switched* connections, where the connection is accomplished manually. The system could be rearranged on request using the same system that transfers regular messages; such requests would be a form of *signaling*.

A third option would be to replace the manual cross-connect by some kind of electromechanical switch and allow the circuits to be built on-the-fly by the introduction of *electrical signaling*. That is, a party in New York wishing to send a message to Los Angeles would have the ability to signal his desire to the system, the system would automatically build the connection from New York to Los Angeles, the party would send its message, and then the system would tear the circuit down. The service is still message delivery, but it is delivered over a temporary circuit-switched connection; the connection lasts only as long as it takes to transmit the message.

A fourth option would be to replace the cross-connects by computers and to terminate each of the channels of the carrier system at the computer. The destination address of each message would be included with the message, and the computer would determine the next hop to which the message would be sent. Thus, as in the manual system, the message would be conveyed hop-by-hop until it reached its destination. As in the first option, this is a message switching system, but now the routing is handled by a computer called a *message switch*.

The introduction of switches into a communication system results in fundamental changes. First, it results in a natural separation of access lines from carrier systems, with the switch being at the interface; the purpose of the access line then is to connect to the switch, and the purpose of the carrier system is to interconnect switches. Second, signaling is required within the system to enable it to respond dynamically to the desires of the end user. A third, less obvious, change is that the introduction of switches enables tremendous flexibility in allocating resources to end users; that is, users can be allocated resources dynamically for short periods of time on demand. This change opens the door for a pay-by-use service policy, a policy that is absent in point-to-point, permanently connected systems.

A switched telecommunication network then has at least the following components: the end user's terminal equipment, switches, local loops between the customer premises and the switches, carrier systems between the switches, and a control system to process the signaling information.

IV. EARLY EVOLUTION OF THE TELEPHONE NETWORK

Voice telecommunications began with the invention of the telephone in 1876. First, we learned to talk between two end points at short distances; end users rented their telephones, but provided their own wires. Next, telephone companies began providing wires, and manual cross-connects were added to allow a calling party to connect to any other party within a local area through the assistance of an operator. Finally, the scattered local switches were connected together in a network, thus allowing subscribers to connect to other subscribers at distant locations.

There is a fundamental difference between a person-to-person voice service and a message-switching service. Unlike message switching, a voice conversation is carried out in real time; the parties at each end of the connection are present and participating simultaneously. This means

that the system must be able to set up a communication path between the end points quickly and also guarantee quality of service for that connection for the duration of the call.

The problem posed by these factors is that a separate circuit is required for each pair of individuals involved in a conversation. Between a person's residence and a switch, the problem is solved by the installation of a local loop. Between two switches, the problem is solved through carrier systems, as described in the previous section.

As the capacity of carrier systems was being improved, there was also a persistent advance in switching technology, including signaling capability. As the signaling capabilities were improved, end customers gained the ability first to dial their own connections within neighborhoods, then within metropolitan areas, then within countries, and finally to anywhere in the world.

The initial carrier systems were based on frequency division multiplexing as described in Section III. In the long-haul part of the network, the frequency division multiplexing system was organized in a hierarchy called the *L-multiplex hierarchy*. At the lowest level, twelve voice circuits were multiplexed together to form a group. Then five groups were multiplexed together to form a *supergroup*. Ten supergroups were combined to form a *mastergroup*. The earliest long-haul transmission systems used coaxial cable, and they carried one mastergroup per cable. Similarly, the microwave radio systems carried one mastergroup per channel. Later, these capacities were raised considerably.

A channel capable of carrying one telephone conversation between two circuit switches is called a *trunk*. The number of trunks between two switches is then equivalent to the maximum number of calls that can be handled simultaneously between the switches. The actual number of trunks installed between switches can be in the hundreds of thousands and is dependent upon the desired rate of calls to be handled by the trunks, the holding time of the calls, and the target quality of service. Quality in this context is measured in terms of the proportion of failed call setup attempts, called *blocking probability*. A tight balance between the number of trunks, placement of switches, and the sizes of the switches is required to achieve high quality at minimal cost. The art of designing a network of switches and trunks is called *traffic engineering*.

The actual placement of multiplexers and switches is determined by cost considerations. For example, suppose that instead of a residence, the telephone service is terminated at a company location that has 1000 employees, all of whom need to use telephone on a regular basis. It would then seem reasonable to develop a class of small switches that can be located at the company premises; such a small switch is called a *branch exchange*, and it may be privately owned, in which case it is called a *private branch exchange*. If this location is at some distance from a telephone switching office and there is a demand for a large number of simultaneous calls, it would also seem reasonable to use a special carrier system, a *loop-multiplex carrier system*, rather than a collection of local loops to interconnect the telephone company's switch and private company's branch exchange.

So far, then, we have the following system components: switches located in strategic places, carrier systems between the switches, local loops between switches and residences, and telephones at the residences. If the switches are far apart, the carrier systems are long-haul, and if the switches are close together, the carrier systems are short-haul.

This model of a network of switches, carrier systems and trunks, and local loops generalizes to almost any conceivable combination in the real world. If one can draw a diagram of any shape, size, or form using these components, then someone has probably built it to meet some need at some time.

The system just described in terms of voice communications provides a service called *circuit switching*. A message switching system could be built over a circuit-switching system; one would simply place message switches wherever one chose and then connect the message switches together using circuits. Therefore, circuit switching can be viewed as a *transmission technology* while voice and message switching can be considered *service technologies*.

Morse's system of dots and dashes for encoding characters and words is related to the system of ones and zeros, called *bits*, that are now familiar in computing systems. Messages consist of strings of characters, characters consist of strings of ones and zeros, and the representation of the message by the string of ones and zeros is referred to as *digital* representation. There would be a strong motivation to invent a machine that could convert digital representations to electrical signals that resemble voice signals so that messages could be transmitted over ordinary voice circuits; such a machine is called a modulator. Once the leap is made to convert data-like signals to voice-like signals, the necessity for a demodulator to perform the reverse process is obvious; otherwise one cannot recover the original signal. The combined machine is called a *modem*. And, once the process of converting between voice-like and data-like signals is well understood, the desirability of converting actual voice signals to digital representation for transmission naturally arises.

A continuous signal, such as voice, in its natural setting is called an *analog signal*, and the process of converting an analog signal to a digital representation is called *analog-to-digital*, or A/D, *conversion*. The combined system that

performs A/D conversion, transmission of digital signals between the end points, and D/A conversion is called a *digital transmission system*.

The question of relative efficiency of analog versus digital transmission of all kinds of signals then arises in a natural way. The first step in settling this question was ushered in with the advent of the digital transmission system in the late 1950s. This first commercially successful digital carrier system was a short-haul carrier system having a capacity of 24 voice circuits per channel, and it was called the *T1 Carrier System*.

Unlike the earlier analog transmission systems, which use *frequency division multiplexing* based on carrier modulation, the T1 Carrier System uses *time-division multiplexing* and transmits bits at baseband. In time-division multiplexing systems, time is organized as a sequence of frames, and frames are divided into slots. Each circuit is allocated one time slot in each frame, and the sequence of slots in successive frames gives the illusion of a full-time circuit, provided that the slots have sufficient capacity.

Analogous to the L-multiplexing hierarchy there is a digital hierarchy, which developed between 1960 and 1980. The lowest level is DS-1, and it carries 24 voice calls operating at 1.5 Mbps. The next level is DS-2, and it carries 96 voice calls and operates at 6.2 Mbps. The next level is DS-3, and it carries 572 voice calls and operates at 45 Mbps. Because T1 carrier systems have the same bit rate as a DS-1 system, lines having the DS-1 rate are sometimes called T1 lines, and similarly, DS-3 lines are sometimes called T3 lines.

By the 1960s, computer concepts were integrated into virtually all parts of circuit-switched networks, and the computer and communications businesses were on a collision course. Switching and carrier systems had been improved to the point where they had enormous capacities, cost had come down significantly and were continuing to fall, and demand for telecommunications services, including data, was increasing.

V. EARLY DATA COMMUNICATION NETWORKS

During the late 1950s, the line between data processing and and telecommunications became increasingly blurred. Telephone companies had significant computer skills resulting both from the fact that telephone switches had long since been controlled by (permanently) stored programs and the fact that expertise had become necessary to manage large businesses. Telephone companies had also realized that once voice is in digital format, one can store it to provide voice mail and also process it in many other ways.

This brought up the issue of whether or not the telephone company could process data commercially, and the resolution of this issue was part of the now well-known consent decree of 1956. In that decree, the Bell System was restricted from entering computer and information services markets.

There are a number of motivations for improving technologies, all of them driven by a desire to lower the cost of providing current services and to increase the possibility of being able to offer new services. In the early 1960s, motivated by the desire on the part of a large number of individuals to access a relatively small number of serious computing resources, the idea of time-sharing was conceived. Time-sharing became commonplace by the late 1960s.

Considerable interest in alternate forms of data transmission technology was evident during the 1960s. L. Kleinrock's pioneering work on *message-switching* systems in the early 1960s is generally acknowledged to be a key catalyst in this movement. One advantage of message-switching systems over circuit-switched connections is *statistical time-division multiplexing*. At each message switch along a route, messages can come in from a number of incoming lines or terminals simultaneously, and it is not always possible that these messages could go out at the same time. The solution is to form a waiting line for those messages that are to be transmitted over the same line, and then transmit them in sequence as time becomes available.

Packet switching, a name coined by D. W. Davies in 1965, is a natural extension of message switching, although message switching is really a service technology and packet switching is a transmission technology. In packet switching, messages are divided into smaller units called *packets* for transmission. Time savings are achieved in multihop connections by transmitting packets in parallel over the various hops of the network.

As an illustration of the gains that can be achieved from packet switching, consider a three-hop connection. If a long message is transmitted as a single unit, then the transmission time will be three times the time required to transmit over one hop. If the message is broken into ten small units, then the total transmission time will be the time required to transmit the entire message over one hop plus the time required to transmit one-tenth of the message over the second and third hops. The reason for the savings is that earlier packets can be transmitted over the second and third hops while later packets are being transmitted over the first hop. Thus the total time required is reduced from three message transmission times to approximately 1.2 transmission times. The relative savings increase when messages are divided into even smaller units up to a certain extent or if the route has more hops.

During the 1960s, engineers and scientists were beginning to gain tremendous insights into the savings in transmission costs and increases in system capacity that could be achieved by digitally processing signals. For example, by including special purpose circuitry at the interface between modems and telephone lines called *equalizers*, more bits per second could be pushed through voice telephone lines. Also, by removing redundancy in data files before transmission, the data transferred per bit transmitted could be maximized.

Significant advances were made in the understanding of data transmission, information theory, coding theory, and signal processing. At the end of the 1960s, many of the advancements in computers and communications were brought together when the world's first packet-switching network, the ARPAnet, came into existence.

As the per-unit cost of electronics decreased in the early 1970s, the proliferation of computer equipment increased. Spurred by the demand for increased services, mainframe divisions and minicomputer divisions within and across company boundaries invented their own techniques for implementing those services. Consulting and maintenance expertise were spread thin throughout the computer industry, and this led major players to try to standardize techniques, at least within their own organizations. The situation was so bad that the very survival of some companies was at stake.

As equipment became more and more distributed within company locations, the demand for sharing peripherals and exchanging information among the computers increased. This led to the development of proprietary means of exchanging data among computers over small areas. The resulting on-campus networks were the earliest examples of *local area networks* (LANs). Numerous LAN protocols were invented, with Ethernet and its derivatives being the most commonly recognized and popular survivor.

The mid-1970s was a period of rapid change in telecommunications. First of all, computer communications was emerging as a field; the idea of development of standard protocols to handle just about any communication need was being formalized. Second, the use of long distance had increased to the point where managing the long-haul system on a heavy calling day, such as Mothers' Day, was no longer possible; the only workable solution was to disallow any alternate routing within the long-distance network during such periods. Third, miniaturization of electronic circuits was progressing, and the first microcomputer appeared.

To simplify software development and facilitate system deployment, communication protocols are organized in *layers*, and the collection of layers is called the *protocol stack*. At the top of the stack is the software that interfaces to the application, and at the bottom is the software that deals with the transmission of bits over a physical facility. A computer program that implements a layer of a protocol stack in a system is called an *entity*. In general, from a service point of view, end systems implement all layers of a protocol stack and intermediate systems implement only the bottom three layers. Layer 3 is called the *packet layer* because it provides a packet transfer service between the layer 4 entities implemented in different end systems. Thus, the packet-switching part of a communications network consists of the interconnection of the layer 3 entities in all the systems of the network.

With miniaturization and integrated manufacturing of electronic circuitry, the mid-1970s saw the advent of the microprocessor. This innovation led to inexpensive processing and resulted in technical revolution in many areas. Among the most revolutionary event was the advent of the personal computer and the accompanying explosion in the number of computers owned by individuals. By the late 1970s, personal computers were already in their second generation, and there were numerous players in the market, led by Apple Computer. User groups, focusing on various personal computers, had begun to emerge and they communicated with each other, mainly by exchanging e-mail over message-switching systems and by posting messages to bulletin boards. Many of these bulletin boards were implemented on personal computers attached to networks via modems.

At the same time, companies were developing products such as data processing platforms, front-end processors, packet switches, retail store controllers, electronic cash registers, and anything else that could conceivably be built around microprocessors. This raised the demand for increased processing power, which was met by an evolution from the 8-bit to the 16-bit processor by the end of the 1970s. The demand for low-cost memory shot through the roof, and this led to enormous efforts to advance memory manufacturing, which at that time seemed to have dead ended at about 16 kilobytes per chip. In addition, special-purpose microprocessors were beginning to appear in the marketplace, and developers began to implement the ideas that had been published in the 1960s.

During the mid-1970s, LANs began to realize success in the marketplace. At this time, their protocols were proprietary, and, with the success of LANs came the demand for standardized products in the LAN area. This demand led to the formation of the IEEE 802 Committee, which developed a set of three standards for LANs by the mid-1980s, one each for *token bus*, *token ring*, and *carrier sense multiple access* (CSMA). While all of these standards were widely deployed, the CSMA version, of which Ethernet is an example, dominated the market by the end of the 1980s.

In the late 1970s and through the 1980s, there were two popular views on how packet-switching networks should be implemented: datagram and virtual circuit. In a datagram service, each individual packet makes its own way across the network, with routing decisions from source to destination being made on each individual packet. Each packet, then, must contain complete addressing information, as in a postal letter, but the switches do not need any state information about the service being rendered. Datagram service was the service of choice for the ARPAnet.

By contrast, a virtual circuit connection uses a call connection phase in which the path that the packets will be taking throughout the duration of the call is set up. Thus, information overhead for each packet can be minimized, but the packet switches must maintain state information on the calls. As mentioned earlier, communication protocols are organized in layers. The three lower layers of the protocol stack that implement this idea are collectively referred to as *X.25*. The third, or highest, layer of X.25 is the packet layer, as mentioned above in the context of general protocols. Layer 2 is called the *frame layer* because layer 2 provides a frame transfer service between the layer 3 entities of two systems. Virtual circuit service was the service of choice of the International Standards Organization (ISO), whose approach was called *Open Systems Interconnection* (OSI).

The 1980s saw continued effort in the networking, personal computer, and workstation areas. A fundamental breakthrough in deployment came in the beginning of 1983 when TCP/IP was deployed throughout the ARPAnet. Although it was not clear at the time, this change clearly delineated the networking element from the end-system element, and ultimately resulted in the IP-network, or Internet. Nodes of the network could now be thought of as coming in two flavors: end systems (hosts) and intermediate systems (routers). That is, the network more or less became a network of routers with the end systems being attachments to the networks or routers.

This separation of functions was extremely important because it partitioned the networking software development effort into manageable portions. End systems ran the TCP protocol, and applications could be built above TCP in the end systems. Both routers and hosts ran IP, and protocols could be developed below IP to provide transmission between routers or between hosts and routers.

At the same time, the *Domain Name System*, which allowed translation between descriptive names for hosts and their Internet addresses, was introduced.

Layers below the IP layer gradually evolved during the 1980s. While X.25 was a very successful protocol, it was somewhat too heavy in that it required packet acknowledgements over each hop of the connection. It was realized that intermediate acknowledgements could be eliminated with very little loss in performance and, instead of involving a third layer, frames could simply be relayed at the second layer. The protocol that implements this idea, *frame relay*, was developed in the late 1980s and deployed in the 1990s. It is used to transfer frames containing packets of any type between packet switches. The net effect is that packet transmission facilities of appropriate capacities between packet switches of networks, especially intranets, can be built economically using frame relay services.

Rapid deployment of computer communication networking products continued throughout the 1980s. Much of this was attributable to two factors: free availability of the Unix operating system on a wide variety of platforms, and free availability of Unix-based networking software. That is, the Unix operating system had become the operating system of choice among software developers in university and research environments, and the advancements in communications protocol development became immediately available to those using the Unix operating system.

Much of the computer communications networking activity of the 1980s focused on building routers to facilitate deployment of networks. Information and people on different networks could be reached if one could figure out the right combination for getting from one network to another. Thick books that gave the relationships among networks became available, and knowledgeable people could move information from a host on one network to a host on a different network; internetworking was beginning to arrive. While computer communications was advancing, the telephone network was not standing still. This topic will be addressed in the next section.

There were a huge number of interconnected networks by the end of the 1980s, mostly deployed in high-tech companies and in academia, and with a significant proportion of total bandwidth in the United States funded by the federal government. Yet, at this time, this technology was largely unavailable to the average person.

From this point, it became natural to move to the point where all networks were considered to be domains of the same network and the domain name system could become ubiquitous. This brought us to the early , when the potential of networking as a commercial enterprise first began to be realized.

VI. COMMON CHANNEL SIGNALING IN THE TELEPHONE NETWORK

During the early 1970s, it was discovered that the overload problem in the long-distance network on heavy calling days was an artifact of the signaling methodology.

Specifically, signaling took place over the same circuits that would eventually carry the call. Constructing a circuit was done sequentially; attempts to find an open circuit would use virtually any available outgoing path. Two unfortunate results were that signaling consumed a disproportionate amount of the trunk resources and the call itself consumed many times the minimum resources required. For example, a call from Miami to Atlanta might consume any number of links, including some to and from the West coast, before finally reaching its destination. Signaling messages themselves also consumed trunking resources so that the effective holding time of the call was increased, and, in addition, signaling might involve numerous trunks that ultimately would not be part of the call, thus adding additional load to the network.

With alternate routing turned off, the situation would be drastically different. For example, a semipermanent route might be set from Miami to Jacksonville and then to Atlanta. If either Miami to Jacksonville or Jacksonville to Atlanta had no available trunks, a call from Miami to Atlanta would be blocked and the caller would be given a busy signal.

A solution to this problem was conceived in the early 1970s and its implementation began in the mid-1970s with the introduction of *common channel interoffice signaling*, or CCIS. Telecommunications engineers had realized that signaling and network management functions could be separated from the call-carrying functions, and separate networks could be designed for each. The call-carrying network would still be circuit switched, but it was decided to implement common channel signaling over a packet-switched network, which was called the *common channel signaling network*. This network was implemented over a protocol stack *signaling system* 7, or SS7.

Thus, from the late-1970s onward, the telephone system became two integrated networks: the SS7 network, which is packet-switched, and the call-carrying network, which is circuit-switched. In order to set up a call, packets are sent around the network, and when it is ascertained that the called party can accept a call and a path of trunks connecting the calling party to the called party is available, the call is connected. No call-carrying capacity is wasted in trying to set up calls for which there were no resources.

By the mid-1980s, new ways of using the SS7 network had emerged. One such use is the introduction of new services, such as 1-800 and call-forwarding, into the telephone system. This was accomplished through enhancements to the control processes of the electronic switches attached to the SS7 network and the introduction of *service control points* attached to that network. The new control software in the switches enabled the switches to communicate with the service control points when necessary to realize the new service. This concept is called the *intelligent network*, and originally, each service control point was dedicated to a specific service. Later the role of service control points was expanded to handle collections of services, and the new concept was called the *advanced intelligent network*.

VII. RECENT DEVELOPMENTS IN TELECOMMUNICATIONS

The U.S. Department of Justice filed an antitrust suit against AT&T in 1974. This suit was settled in 1982 leaving a fragmented U.S. telecommunications business in place in 1984. In the meantime, many major problems standing in the way of the deployment of fiber-optic communications systems had been solved starting in the mid-1970s, and significant advancements had occurred in fiber-optic technology by 1984. These advancements made it possible for more companies to deploy fiber-optic systems at a reasonable cost, and, hence, from a technological point of view, opened the way for more competition in the long-distance market.

Other innovations of the 1980s include the widespread availability of community antenna television (CATV), or cable television, an increase in the use of communication satellites, and the advent of cellular telephony. Cable television, begun as a niche business in the late 1940s to provide television services in remote or mountainous regions, had evolved a new role, where it facilitated the delivery of a wide choice of television programming to its subscribers. Deployment of CATV continued throughout the 1980s and 1990s, and, in addition, satellite-based television was widely deployed. In the latter half of the 1990s, satellite-based digital TV with 18-inch receiving antennas was deployed.

As the cost of electronics and electronic systems plummeted during the 1970s, demand for mobile communications, especially telephone, increased. Because additional frequency spectrum to support the increased demand was not available, techniques for using the existing spectrum more efficiently were aggressively sought. With the improvement in networking technology and electronic switching, the concept of cellular mobile telephony soon developed.

In cellular telephony, a metropolitan area is partitioned into a grid of smaller areas called *cells*. At the center of each cell is a *cell site*, which consists of antennas, transmitters and receivers, and miscellaneous control systems. Each cell is allocated a subset of the available frequency channels, with multiple cell sites within the same metropolitan area having the same sets of frequencies. The

system is designed so that transmissions from cell sites having the same collection of frequencies do not interfere with each other; cell sites having the same frequencies are farther apart than those having different sets of frequencies, and, in addition, transmitter power levels are carefully controlled.

The system just described uses analog transmission and frequency division multiplexing. Other systems using digital transmission were developed during the 1990s. One of these digital systems uses a combination of frequency and time-division multiplexing and the other uses a combination of frequency division multiplexing and *spread spectrum multiple access*. These new technological innovations have raised both the quality and the voice call-carrying capacity of cellular telephony. In addition, these innovations have lead to serious research in the area of wireless data access.

An advantage of wireless cellular access over traditionally wired local access is that it can be deployed to an area quickly. Specifically, once a telephone switch is set up in a local area, a cellular telephony system can be deployed, and telephone service can be provided in the local area without installing any local loops. After a long-distance infrastructure is deployed, worldwide telephone service can be realized almost overnight. Because of these properties, cellular telephony is finding a market in countries where very little traditional telephone infrastructure exists today.

The late 1980s and early 1990s also witnessed continuing advances in fiber-optic technology. With the rapid development of optical fiber-based systems, a third multiplexing hierarchy, the *synchronous digital hierarchy* or SDH, was developed. The lowest level of the SDH is OC-1, which has a capacity of about 50 megabits/second and carries the equivalent of 672 voice calls. Popular levels of SDH include OC-3 at about 150 megabits/second, OC-12 at about 600 megabits/second, OC-48 at about 2300 megabits/second, and OC-192 at about 9600 megabits/second. At least one commercial packet switch that can handle OC-192 lines exists today, and the potential to develop packet switches that can handle line rates of OC-768, the equivalent of 500,000 voice calls, is a frequent topic of discussion in research circles. Today, one can buy off-the-shelf equipment to demultiplex an OC-192 stream into 16 OC-12 streams.

During the 1990s, a number of technological refinements of and extensions to basic internetworking capabilities were introduced. Among these were 100 megabit/second Ethernet, wireless LANs, Ethernet hubs and switches, and gigabit/second Ethernet.

But the big news of the 1990s was the widespread deployment and commercialization of the Internet. One driving force was the advent of the *World Wide Web*, or just Web. The Web consists of a collection of server locations called web sites, which are accessed over TCP/IP connections using a web browser, the two most popular of which are Netscape's Navigator and Microsoft's Internet Explorer. *Hypertext mark-up language* (HTML) is used to format web sites, and the *hypertext transfer protocol* (HTTP) is used for interaction between clients and servers, or equivalently, between web sites and web browsers.

Because a web browser's interface is graphical and intuitive, anyone can learn to use the Web quickly. This has led to the creation of many new companies, the development of many web sites, and the expansion of networking services into many millions of homes worldwide. Thus, in the 1990s, packet-based networking services spread from the academic and research environments into the world's businesses and homes, becoming a common household utility on par with the telephone and television.

With the widespread availability of packet-based services, new developments began to occur to provide higher speed data access from the home. Part of this expansion took place with *Integrated Services Digital Network* (ISDN) service, which provides integrated voice and digital services at approximately 144 kilobits/second. This technology developed in the 1980s but did not see broad acceptance in the market until higher speed data access was needed.

The use of SS7 in controlling and managing circuit-switched networks makes it necessary for telephone companies to maintain leading edge competence in both circuit- and packet-switching technology. One consequence of this expertise was the proposal of *asynchronous transfer mode* (ATM), as a solution for integrating multimedia communications using a packet-based technology.

Initially, it was thought that the ATM protocol suite would span the entire range of services from application to application. That is, ATM would be deployed in end systems, in intermediate systems, and everywhere in between. Hundreds of companies worked to develop ATM standards at all layers, and many good ideas came out of those efforts. However, probably because of the broad deployment of TCP/IP based applications and continuing advancements in Ethernet technologies in the local area, ATM has been more successful as a transmission technology than as an end-to-end service technology. Indeed, much of the world's IP datagram transmission takes place over frame relay systems implemented over ATM.

Additional high-speed access from the home was offered through a cable-modem service built over cable television facilities. Still more digital capacity to the home was offered by *Asymmetric Digital Subscriber Service* (ADSL), which uses ordinary twisted pair telephone cable, and which became widely available at the end of the 1990s. As the year 2001 gets under way, there is a continuing effort to increase the rate at which data can enter

the home, especially using technologies that can be implemented over either the existing telephone wires or cable television facilities.

As the technologies already discussed were introduced, improved, or supplanted, the firms founded to develop and market these technologies also changed. Before discussing future trends, we present a brief overview of the evolution of the telegraph and telephone industries in the United States from 1837 to mid-February, 2001.

VIII. EVOLUTION OF THE UNITED STATES TELECOMMUNICATIONS INDUSTRY: 1837–2000

The electric telegraph, a system of communication employing electrical apparatus to transmit and receive information using a code of electrical pulses, was invented in 1837 by Samuel Finely Bresse Morse, an American artist and inventor. In Great Britain in the same year, the British physicist Sir Charles Wheatstone, in collaboration with the British engineer Sir William F. Cooke, also invented an electric telegraph. However, Morse was the first to construct an experimental telegraph line and on May 24, 1844, Morse sent the first message between Washington, D.C., and Baltimore, MD. Morse filed for a patent on his invention in 1838. The patent was granted in 1848.

In 1851, the New York and Mississippi Valley Printing Telegraph Company was formed to build a telegraph line from Buffalo, NY, to St. Louis, Mo. In 1856, the company was reorganized and renamed the Western Union Telegraph Company. By the end of 1861, Western Union had built the first transcontinental telegraph line and there were 2250 telegraph offices in operation nationwide.

In 1854 Cyrus West Field, an American merchant and financier, obtained a charter giving him the exclusive right for the next 50 years to use the coast of Newfoundland for a transoceanic cable terminal. In collaboration with American manufacturer Peter Cooper and others, he formed two companies to lay a transatlantic cable: The New York, Newfoundland and London Telegraph Company and The Atlantic Telegraph Company, The British and United States governments helped fund Field's efforts. In July 1866, the first transatlantic cable was complete.

The telegraph was a very successful technology and Western Union's business grew rapidly throughout the early 1900s. By 1943, after acquiring Postal Telegraph and some 500 other competitors, it was the largest company in its field. However, telegraph rates rose over time while the rates of its major competitor, the telephone, declined. During the 1960s, the company moved into a number of other related services, such as time-sharing computer systems, teleprinters, and satellite communications. Western Union Telegraph Co. was reorganized in 1988. The reorganized firm, Western Union Corp., focused on money transfers and other financial services. In 1991, parts of the company were sold and the company's name was changed to New Valley Corp. New Valley Corp. entered bankruptcy proceedings in 1993 and sold the last of its major holdings, Western Union Financial Services Inc., to First Financial Management Corp.

Alexander Graham Bell, an inventor and teacher of the deaf, was born in Scotland in 1847. Bell immigrated to Canada in 1870 and then to the United States in 1871. Bell was working on a multiple telegraph when he developed the basic ideas for the telephone in 1874. On March 10, 1876, he successfully transmitted the first telephone message. Bell patented the telephone in 1876 and, facing financial difficulty, he offered to sell his patent to Western Union Telegraph Company for $100,000. Western Union declined. In 1877, Bell and three associates, Thomas Watson, Thomas Sanders, and Gardiner Hubbard formed the Bell Telephone Company as a voluntary association. In 1878, the company was incorporated.

By 1877 there were four basic Bell patents. These initial patents protected telephone instruments, not wires or switches, and originally the company rented instruments to individuals who were responsible for providing their own connecting wire. This proved to be an inefficient method for connecting a large number of individuals so subscriber service was initiated. Subscribers paid to be connected to one another through equipment provided entirely by the various telephone companies. Calls were placed through manual exchanges. When a subscriber wanted service they picked up their handset and a small light on a switchboard alerted an operator that the subscriber wanted service. The operator then inserted a cable into the jack on the switchboard that corresponded to the caller. The caller told the operator the called party's name. The operator used another cord adjacent to the first to plug into the called party's jack and then operated an electrical switch that connected ringing current to the called party's telephone. When the called party answered, the operator disconnected leaving the caller and the called party connected. The first commercial exchange, which came into service in New Haven, CT in 1878, connected 21 subscribers. Although these initial exchanges provided only intracity service, telephone service expanded rapidly. There was one exchange with 21 customers in 1878 and there were 400 exchanges with 140 subscribers in 1885. Telephones per capita increased as well. In 1880 there were 1.1 telephones per thousand and by 1894 there were 4.1 telephones per thousand.

In 1878, Western Union Telegraph and the Bell corporation both operated telephone and telegraph services. The Bell Company filed a patent infringement suit against Western Union and won the suit in 1897. As a result of the settlement, Western Union sold its 55-city telephone system to Bell Telephone and agreed to leave the telephone business while Bell Telephone agreed to leave the telegraph business and pay Western union 20% of their telephone rentals for the life of the Bell patents.

In 1882, Bell Telephone acquired Western Electric Company, the largest manufacturer of electrical equipment in the United States, from Western Union. Bell Telephone's long-distance service was expanding and in 1885, Bell Telephone created a subsidiary, the American Telephone and Telegraph Company (AT&T), to finance, build, and operate the long-distance system. In 1899, the company was reorganized, and AT&T became the parent company for the entire Bell system.

Bell's fundamental equipment patents expired in 1893 and 1894. This led to the entry of competitors, first in areas not served by Bell Telephone, and then in major cities where they were in direct competition with Bell. In 1907, a group of investors led by American financier J. P. Morgan took control of the company and Morgan appointed Theodore Vail president. Vail, in an effort to restore AT&T's market share, began an aggressive program of patent control, buyouts of independent companies, and refusal to allow competitors to connect to AT&T's long-distance network. In 1909, AT&T purchased Western Union. This acquisition, combined with Vail's earlier actions, caused AT&T's competitors to charge that it had become a trust, and the Department of Justice (DOJ) threatened antitrust action. In response to this threat, AT&T entered into an agreement with the DOJ known as the Kingsbury Commitment. AT&T agreed to sell Western Union and to allow independent local telephone companies to connect to its long-distance lines. AT&T and the independents exchanged local territories creating effective monopolies in each local area. Many small towns and rural areas in the United States were and are still served by local companies that have never been part of the Bell System (e.g., in 1982, Bell system companies served 81% of the telephone lines but only 41% of the territory in the United States). AT&T, however, maintained its monopoly on long-distance service and retained control of Western Electric. The result of this regulatory structure was to transform the industry from one in which there was competition in the local areas with a monopoly in long-distance service to a fully monopolized regulated industry.

The Bell system was subject to both state and federal regulation. Regulators approved pricing structures, called tariffs, partially based on the cost of providing the service. In a telephone system, parts of the system will be used for both local and long-distance service. State regulatory commissions typically wanted to shift a portion of the cost of local service to long-distance service. This cost allocation process, in which the cost of commonly used plant is divided between local and interstate jurisdictions, is termed *separations*. AT&T opposed this practice. In 1930, the U.S. Supreme Court ruled in *Smith v. Illinois Bell* that some of the cost of the local network had to be allocated to the interstate service. In 1934, the Federal Communications Commission (FCC) was established by the Communications Act. The Act established the Joint Board. This board was composed of federal and state commissioners and was responsible for analyzing the separation of costs and making recommendations to the FCC. In 1947, the Joint Board produced the first separations manual.

In 1949, the DOJ filed an antitrust suit against Western Electric alleging monopolization of the telephone equipment market. In 1956, the DOJ sought to force AT&T to divest itself of Western Electric and to break Western Electric into three companies. The suit was settled by a consent decree in 1956 generally considered to have favored AT&T. Under the decree, the DOJ recognized that the structure of AT&T was not an illegal anticompetitive arrangement and AT&T was allowed to keep Western Electric as a separate subsidiary. However, the decree limited Western Electric to manufacturing equipment for the Bell System and to contract work for the government. Western Electric sold its small nontelephone subsidiary Westrex to Litton Industries and its holdings in Northern Electric (now known as Nortel Networks) to the public. In addition, AT&T was restricted to providing common carrier communications services. This meant that the company was restricted from entering the computer and information services markets.

During the 1950s, there was an increase in applications to the FCC for permission to build private microwave transmission systems. In 1956, the FCC began considering policies to allocate frequencies above 890 MHz. AT&T, Western Union, and the private telephone companies all argued that the use of these frequencies should be limited. Specifically, they did not want private microwave systems to operate in frequencies where common carrier service was available and they argued that the use of the frequencies should be abandoned if common carrier service became available. In 1959, in the Above 890 Decision the FCC ruled that the frequencies could be used by both common carriers and licensed private users but private users were limited to building systems for their own use. In response to this decision, AT&T filed a new tariff, Telpak, that offered large discounts for groups of private

lines. This tariff substantially decreased the incentive to build private microwave systems. As was usual, Telpak prohibited resale and sharing.

However, the Telpak tariff also indicated a potential profit opportunity. A company might identify users that were too small to use the Telpak discounts, set up a microwave transmission system, and share it among those users. In 1963, Microwave Communications, Inc. (MCI) decided to move into this market. MCI filed an application with the FCC to set up a limited microwave transmission system between St. Louis and Chicago. This system would be a public system because its purpose was to sell service so the application could not be handled under the 890 Decision. Again AT&T, Western Union, and the local carriers opposed the request. In 1969, the FCC decided in favor of MCI. AT&T appealed the decision. The FCC then moved beyond the 1969 decision to a fundamental policy change and in 1971 the FCC announced the Specialized Common Carrier (SCC) decision. Under this decision individual licenses for microwave routes were still required but the decision allowed entry by firms that wanted to begin public microwave transmission service. In addition, the FCC ordered that the established common carriers be required to interconnect with any new carriers.

MCI began service between St. Louis and Chicago in January, 1972. The company expanded its service and by 1973 had a basic network to many major cities. In 1973, MCI filed a complaint with the FCC alleging that AT&T was, in general, refusing to allow interconnection and that when AT&T did allow interconnection the terms were not as favorable as those given to Western Union. Between 1973 and 1978, MCI aggressively attempted to expand its service offering beyond simple public line to foreign exchange (FX) and long-distance calling, i.e., Message Telecommunications Service (MTS). AT&T responded with tariffs and appeals of MCI's interconnection requests. In 1978, a settlement was reached that set up a structure for payments between AT&T and its competitors. The FCC also issued a decision invalidating AT&T's restrictions on resale of its services. Switched service competition was established and other long-distance rivals such as United Telecommunications (United Telecom) began to enter the market.

In 1974, both MCI and the DOJ filed antitrust suits against AT&T. The DOJ antitrust went forward and after considerable delay went to trial in 1981. In January, 1982 the suit was settled by an agreement known as the Modified Final Judgment (MFJ). Under the MFJ, AT&T was required to divest itself of its local exchange telephone service. AT&T was allowed to retain Western Electric and Bell Laboratories.

The process of breaking up AT&T took two years; formal divestiture took place on January 1, 1984, when the old Bell System was replaced by 22 Bell Operating Companies, (BOCs), 7 Regional Bell Operating Companies (RBOCs,) and AT&T. The RBOCs, commonly called Baby Bells, were Bell Atlantic, BellSouth, NYNEX, Ameritech, U.S. West, Southwestern Bell, and Pacific Telesis. AT&T's Central Services Organization became Bell Communications Research Inc. (Bellcore). Bellcore's original purpose was to provide technological expertise and research to the Bell Operating Companies. Under the MFJ the seven regional companies were made responsible for all local calling, some intrastate long-distance business, customer access to long-distance networks, and directory advertising. The RBOCs could not compete with each other in the local calling markets. The RBOCs were also restrained from entering the interstate long-distance business. They could, with the permission of the court, enter other businesses.

AT&T's post divestiture business consisted of long-distance services, services for all customer-terminal equipment then in place, research and development, and the Western Electric manufacturing company. Western Electric's charter was assumed by a new unit, AT&T Technologies. After divestiture, AT&T long-distance no longer subsidized local service and AT&T cut prices. MCI was forced to do the same and lost $448 million in 1986. It is important to note that while the MFJ forced the divestiture of AT&T it did not fully deregulate the telecommunications industry. Federal and state regulation remained pervasive and was applied to the RBOCs and to AT&T's competitive telecommunications services but not to AT&T's long-distance rivals.

While divestiture did not fully deregulate the United States telecommunications market, it did change the regulatory climate and open up opportunities. In addition, technological changes were quickly blurring the distinction between telephone networks, computer networks, and broadcast networks. These factors caused new entry as well as merger and acquisition activity to increase between 1983 and 1995. Then in 1996 the U.S. Congress passed the Telecommunications Act which effectively deregulated the United States telecommunications industry. The Telecommunications Act allowed the RBOCs to enter the long-distance market and it allowed the various long-distance providers to provide local service. This resulted in further changes in firm structure. The remainder of this section describes the effect of these changes on some of the major firms in the United States Telecommunications industry.

A. Sprint

In the early days of the railroads, many railroads had installed telegraph wires along the tracks and had developed

long-distance private communications networks using these wires. Southern Pacific's switched private network telecommunications system was one of these wire networks. In 1983, GTE bought Southern Pacific's network and began replacing the copper wires with fiber-optic cables. In 1985 United Telecom, which had its own fiber-optic network, bought 30% of GTE's fiber network. In 1986, the two fiber-optic networks were combined, and a company to manage this network was formed and named U.S. Sprint. In 1989, United Telecom bought the controlling interest in U.S. Sprint from GTE and then in 1992, bought out the remaining portion of GTE's ownership. United Telecom then changed its name to Sprint Corporation.

In 1993, Sprint merged with Centel Corporation, a company specializing in local and cellular telephone services. Sprint then formed an alliance with Call-Net to market telecommunications services in Canada under the name Sprint Canada. In 1994, Sprint signed an agreement with Telefonos de Mexico to cooperate in providing communications services between the United States and Mexico. Also in 1994, the FCC began a series of auctions (still ongoing in 2001) to allocate frequencies. The next year, Sprint formed a joint venture with three major cable TV companies, Tele-Communications, Inc. (TCI), Comcast Corp., and Cox Cable, to buy the rights to Personal Communications Services (PCS) wireless licenses in 29 major United States markets. Also in 1995, Sprint entered an alliance with Deutsche Telecom of Germany and France Telecom. The three firms launched Global One to compete in the international corporate telecommunications market. In 1999, MCI WorldCom proposed a friendly acquisition of Sprint but the acquisition was blocked by the United States DOJ. In February of 2001, Deutsche Telekom announced its intention to sell its 10% stake in Sprint.

B. MCI

In 1993, MCI and British Telecommunications (BT) entered into strategic alliance and BT invested $4.3 billion to acquire a 20% stake in MCI. The next year, MCI and Grupo Financiero Banamex-Accival formed a joint venture known as Avantel to offer long-distance service in Mexico. Competition in the telecommunications industry intensified and MCI lost more than a million customers to AT&T in 1994. In 1995, MCI acquired Nationwide Cellular, an American cellular telephone company, and SHL Systemhouse, a Canadian firm specializing in corporate computer networking systems. In 1996, MCI and News Corporation formed a joint venture to offer consumers information and entertainment through a satellite system, American Sky Broadcasting (ASkyB). Similar satellite services were offered to businesses through the SkyMCI program. MCI also formed strategic alliances with Microsoft Corporation, Digital Equipment Corporation, and Intel Corporation. Later in 1996, BT purchased the remaining 80% of MCI for $20.8 billion. At the time this was the biggest foreign takeover of a United States company in history. In 1998, WorldCom Inc. purchased MCI Communications Inc. from BT for $34.7 billion.

C. WorldCom

Long Distance Discount Service (LDDS) was created in 1983 to resell AT&T long-distance service to small- and medium-size businesses in southern Mississippi. LItDS began an aggressive acquisitions strategy in 1986 purchasing small carriers in MS and TN. In 1993, LDDS acquired Metromedia Communications and Resurgens Communications Group. These acquistions extended its coverage to the entire country. The next year LDDS acquired IDB WorldCom, the fourth-ranked carrier of international long-distance calls. In early 1995, the company acquired WilTel, which owned a nationwide fiber-optic network. After the WilTel purchase, LDDS changed its name to WorldCom, Inc. In 1996, WorldCom bought MFS Communications for $12 billion. Then in 1998, WorldCom Inc. acquired MCI Communications Inc. for $34.7 billion. The combined company was renamed MCI WorldCom. In 1999, in the largest proposed takeover in United States history, MCI WorldCom announced a $115 billion acquisition of Sprint. This acquisition was blocked by the United States DOJ.

D. AT&T

In 1991, AT&T acquired NCR Corp. In 1994, the company acquired McCaw Cellular Communications of Kirkland, WA, then the largest provider of cellular phone service in the United States. McCaw Cellular was renamed AT&T Wireless Services.

In September, 1995, AT&T announced a restructuring plan. The plan divided AT&T into three separate companies: a systems and equipment company, a computer company, and a communications services company. The systems and equipment company was formed by combining AT&T's systems and technology unit and Bell Laboratories. This company was renamed Lucent Technologies. Bell Labs became the research and development arm of Lucent. As a division of AT&T, its scientists had invented the transistor, laser, solar cell, light-emitting diode (LED), cellular radio, C programming language, UNIX operating system, and many other technologies. Lucent Technologies was spun off in September, 1996. The

second company consisted of the unit formed when AT&T acquired NCR in 1991. This company was spun off as NCR Corp. in January, 1997. The third company retained the AT&T name. The "new" AT&T's focus was on becoming an integrated voice and data communications company. The Bell Labs researchers who supported AT&T's pre-reorganization communications services business stayed with AT&T as the staff of the new AT&T Labs.

In 1998, AT&T acquired Teleport Communications Group (TCG). In early 1999, AT&T acquired IBM Global Networks and later that year AT&T merged with the second largest cable company in the United States, Tele-Communications Inc. (TCI). TCI was renamed AT&T Broadband and Internet Services.

In January of 2000, AT&T and British Telecom formed a joint venture, Concert. In late 2000, AT&T announced a new restructuring plan. Under this plan, which is expected to be fully implemented during 2002, AT&T will create four separate, publicly traded securities for its major business units, Broadband and Internet Service, Wireless Services, Business Services, and Consumer Services. After the restructuring is complete, AT&T Business, AT&T Wireless, and AT&T Broadband will be spun off as separate businesses while the consumer services unit will remain a part of AT&T.

E. RBOCs and Bellcore

After passage of the Telecommunications Act of 1996, two Baby Bells, Southwestern Bell and Pacific Telesis, merged in a $16.7 million deal and NYNEX bought Bell Atlantic for $22.1 billion, leaving only 5 RBOCS. In 2000, Bell Atlantic merged with GTE. As part of the merger GTE named its Internet unit Genuity and that unit was spun off.

In 1997, Bellcore was purchased by Science Application International Corporation (SAIC), and in 1999 Bellcore's name was changed to Telcordia Technologies.

IX. THE TWENTY-FIRST CENTURY

With the increased availability of IP networking services to the home, not to mention businesses, the question "Why not IP Telephony?" naturally arises. Voice over IP, or VoIP, does exist today, and many believe that VoIP will provide most voice services before the year 2010. There are many problems to solve, primarily in the area of guaranteeing quality of service, but progress is steadily being made to define specific problems that must be solved.

We have mentioned briefly that one of the main advantages of packet-switching transmission infrastructure is statistical multiplexing gain, which allows bandwidth sharing among users of the same services and across services. Many researchers view this as a prime motivation for implementing all services on packet-switched networks. However, using packet-switching as an integration strategy requires the invention and deployment of strategies that guarantee adequate quality of service in a heterogeneous environment. History tells us that people simply will not be happy with marginal voice quality; this has been demonstrated time and time again through experimental deployment of systems that provide lower voice quality for lower prices. This situation is exacerbated in multiparty teleconferencing.

Some argue that with the ability to set up circuit-switched connections blindingly fast and the advent of virtually infinite bandwidth through the worldwide deployment of fiber-optic based transmission systems, the natural evolution will be toward an all circuit-switched based infrastructure.

On the surface, an all circuit-switched approach seems to make some sense, but it has at least one major pitfall, which is that it inherently assumes allocation of specific pieces of equipment to a single service. Such an approach has an inherent limitation in that it limits the number of simultaneous services that can be provided to the same place at the same time to a fixed number. Even at this time, the limitations of such an approach as a long-term solution are obvious.

A possible solution to this dilemma is to provide circuit-switched services over a packet-based infrastructure. Such a solution is easily developed using the concept of virtual circuits and its variants. The concept of virtual circuits has been around since at least the mid-1970s, and one of its variants, *label switching*, is beginning to gain momentum today. In addition, much insight into the problems that must be solved and how to solve them was developed during the mid-1990s in the context of ATM.

The first ten years of the twenty-first century promise to be an exciting time for telecommunications from both technology and service deployment points of view.

X. TELECOMMUNICATIONS TIMELINE: TELEGRAPH, TELEPHONE, TELEVISION, COMPUTER, INTERNET: 1837–2001

1837 Morse invents telegraph.
1938 Morse files patent on telegraph.
1839 Becquerel discovers the electrochemical effects of light.
1844 Morse sends first telegraphic message.

1847 Blakewell patents the chemical telegraph.
1851 New York and Mississippi Valley Printing Telegraph Company is formed.
1854 Field forms two companies to lay transoceanic telegraph cable.
1856 New York and Mississippi Valley is reorganized and renamed Western Union Telegraph Company.
1861 Western Union builds first transcontinental telegraph line.
Caselli uses tin foil on facsimile to transmit handwriting and pictures.
2250 telegraph offices in the United States.
1866 Field's companies complete the first transatlantic telegraph cable.
1870 Edison invents multiplex telegraphy.
1876 Bell transmits first telephone message.
1877 Bell, Watson, Sanders, and Hubbard form the Bell Telephone Company as a voluntary association.
1878 Bell Telephone Company is incorporated.
First commercial telephone exchange in operation in New Haven, CT, with 21 subscribers.
1879 Western Union sells its 55 exchange system to Bell and Bell agrees to leave the telegraph business.
1882 Bell Company acquires Western Electric.
1883 Edison discovers the Edison effect.
1885 AT&T is created to finance, build, and operate Bell Telephone's long-distance service.
1890 Hollerith's method of using punched cards to tabulate data is used in United States census.
1893 Fundamental Bell patents expire in 1893 and 1894.
1896 Hollerith starts the Tabulating Machine Co.
1897 Brun invents the Cathode Ray Tube (CTR).
1899 AT&T becomes parent company for the Bell System.
Marconi's Wireless Telegraph Company, Ltd. establishes communications across English Channel.
1900 Approximately 20,000 telephone operating companies (Telcos) and 856,000 telephones in the United States.
1906 deForest invents the vacuum tube.
1907 J. P. Morgan group takes control of AT&T. Vail appointed president.
1909 AT&T purchases Western Union.
Marconi and Braun are awarded Nobel Prize for work in wireless telegraphy.
1910 Interstate Commerce Commission (ICC) given authority to regulate interstate phone service.
1911 The Tabulating Machine Co. merges with three other firms to form the Computing-Tabulating Recording Co.
1913 Bell System and the DOJ reach agreement with regard to interconnection; the Kingsbury Commitment.
1919 Westinghouse, AT&T, and United Fruit form RCA which operates the American Broadcasting Company (ABC).
1920 Valensi develops the concept of time domain multiplexing.
1923 Zworykin patents an electronic camera tube, the "ionoscope."
Westinghouse, General Electric, and AT&T begin television research.
1924 The Computing-Tabulating Recording Co. changes name to International Business Machines (IBM).
1925 Bell Telephone Laboratories is founded.
Bush at MIT builds a large-scale analog calculator; the differential analyzer.
12 million phones in service in the United States.
1926 Pulse Code Modulation (PCM) is developed by Rainey at Western Electric.
RCA creates the National Broadcasting Company (NBC).
1927 AT&T performs first long-distance TV broadcast from Washington, D.C. to New York.
RCA creates Columbia Broadcasting System (CBS).
1928 First television station starts broadcasting in Schenectady, NY.
1930 In *Smith v. Illinois*, a portion of the cost of local telephone service is allocated to long-distance.
Farnsworth patents electronic television.
1932 RCA's control of ABC and CBS is declared a violation of antitrust laws.
1934 Federal Commerce Commission (FCC) is established.
1935 First telephone call around the world.
Approximately 6700 Telcos in the United States.
1936 The first public broadcasting of television programs takes place in London.
Turing publishes "On Computable Numbers."
Zuse applies for patent on mechanical memory.
1938 AT&T introduces crossbar central office switches.
Zuse creates the Z1.
1939 Atanasoff and Berry at Iowa State University invent the first electronic digital computer, the Antanasoff-Berry Computer.
Stibtz at Bell Labs completes the Complex Number Calculator.
Hewlett and Packard found Hewlett-Packard (HP).
1941 Zuse develops the first programmable calculator, the Z3, using binary numbers and Boolean logic.
NBC and CBS are granted commercial television broadcast licenses.

The United States adopts a 525-line black and white system as the standard for TV broadcasting.

1943 Western Union and Postal Telegraph merge.
RCA sells ABC to Noble.

1944 Harvard-IBM MARK I, a large-scale calculating machine, is used in WWII by the U.S. Navy.
The Colossus, a programmable digital machine, is used to break German codes.

1945 Electronic Discrete Variable Automatic Computer (EDVAC) is created.

1946 Raytheon transmits audio via microwave link from WQXR in New York City to Boston.
Turing designs the Automatic Computing Engine.
Williams applies for patent on CRT storage device.

1947 Bardeen, Brattain, and Shockley invent the germanium point contact transistor.
Harvard MARK II goes into operation.

1948 Bell Telephone Laboratories introduces the alloy junction germanium transistor.
CBS announces development of color television.
IBM builds the Selective Sequence Electronic Calculator.

1949 Shannon of Bell Lab's publishes his seminal theory of relay logic.
Joint Board produces first separations manual.
DOJ files antitrust suit against Western Electric.

1950 Remington Rand buys Eckert-Mauchly Computer Corporation. Name is changed to UNIVAC Division of Remington Rand.

1951 The UNIVAC is the first computer to be sold commercially in the United States.

1952 The first database is implemented on RCA's Bizmac computer.
Remington Rand buys Electronic Research Associates and combines it with the UNIVAC Division.
More than 22 million TV sets in the United States.

1953 ABC merges with Paramount Theaters.

1954 Amdahl develops the first computer operating system for the IBM 704.
IBM introduces Remote Job Entry (RJE).
Sony introduces the first transistor radio.
Commercial color broadcasting begins in the United States using NTSC standards.
IBM produces and markets the IBM 650.
IBM publishes the first version of FORTRAN.
Teal at Texas Instruments (TI) discovers how to make a transistor out of silicon.

1956 AT&T enters into a consent decree with the DOJ to settle the 1949 antitrust suit.
Bell Labs scientists Bardeen, Brattain, and Shockley share the Nobel Prize in physics.
The IBM 305 RAMAC is shipped.

1957 FCC Hushaphone Decision allows foreign equipment to be attached to Bell equipment.
Russia launches the first satellite, Sputnik.
Digital Equipment Corp. (DEC) is founded.
Philco produces the TRANSAC S-2000.

1958 Scientists at Fairchild and TI developing the first integrated circuits.
TI introduces the silicon-based transistor.
Cray at Control Data Corporation (CDC) develops the first transistorized computer, Model 1604.

1959 Above 890 Decision: FCC approves private microwave communication systems.
AT&T introduces the Telpak tariff.
AT&T introduces the TH-1 1860-channel microwave system.

1960 AT&T installs the first electronic switching system in Morris, IL.
COBOL is invented.
IBM producing the IBM 1400.
3299 Telecos in the United States.

1962 AT&T introduces T-1 multiplex service in Skokie, IL.
Baron of RAND introduces the idea of distributed packet-switching networks.
Comsat is formed.
ABC requests FCC to allow domestic satellites to distribute TV programs.
Telestar 1 satellite launched.
Fairchild Semiconductor and TI begin mass-producing the integrated circuit.
Approximately 10,000 computers in service in the United States.

1963 Microwave Communications Inc. (MCI) files an application with the FCC to set up a public microwave system between St. Louis and Chicago.
Englebart at SRI patents the idea of the mouse.
The American Standard Code for Information Interchange (ASCII) is developed.

1964 IBM releases the Model 360 computer.
American Airlines and IBM introduce the Semi-Automated Business Research Environment (SABRE) system to computerize airline reservation.
Heilmeier at RCA invents the liquid crystal display.
Kennedy and Kurtz at Dartmouth University develop Beginner's Allpurpose Symbolic Instruction Language (BASIC).

1965 AT&T introduces stored program controlled switching.
Commercial satellite Early bird launched in fixed orbit.

Donald Davies coins the term "packet switching."
Marill and Roberts set up first Wide Area Network (WAN) between MIT's Lincoln Lab and System Development Corporation.
2421 Telcos in the United States.

1968 Carterfone Decision; FCC rules that AT&T has to allow devices that do not cause harm to the network.
AT&T starts 56 kilobits/second service.
Noble at IBM develops the 8-inch floppy disk.
Kramer at Phillips invents the compact disk.
Noyce and Moore co-found Intel Corp.

1969 Bell Labs develops UNIX.
Department of Defense (DOD) sets up Advanced Research Projects Agency (ARPAnet) and the first four ARPAnet nodes are installed; IMP1 at UCLA, IMP2 at Stanford Research Institute, IMP3 at UC Santa Barbara, and IMP4 at the University of Utah.
Starkweather at Xerox invents the laser printer.
Control Data Corp releases the CDC 7600, the first supercomputer.
CompuServe, the first commercial online service is established.

1970 AT&T introduces the 2ESS electronic switch.
Intel introduces the 4004 4-bit microprocessor and the 1103 memory chip.
1841 Telcos in the United States.

1971 Specialized Common Carriers (SCC) Decision; FCC rules that MCI and others may provide public microwave service.
Intel designs the first microprocessor, the Intel 4004.
Wirth invents Pascal.
ARPAnet has 15 sites and averages 700,000 packets a day.

1972 MCI begins public microwave service between St. Louis and Chicago.
Ritchie at Bell Labs invents C.

1973 MCI files interconnection complaint with FCC.
The File Transfer Protocol (FTP) is introduced.
Metcalf creates Ethernet, a local area network (LAN).

1974 MCI files private antitrust suit against AT&T.
DOJ files antitrust suit against AT&T.
IBM announces System Network Architecture (SNA).
First domestic satellites are in operation.
AT&T introduces the digital subscriber loop (DSL).
Cerf and Kahn create Transmission Control Protocol (TCP).
Kilty, Merryman, and Tassel of TI patent the electronic hand-held calculator.
Intel 8080 processor becomes the industry standard.

1975 Micro Instruments Telemetry Systems ships the Altair 8800.
Bill Gates and Paul Allen start Microsoft.
1618 Telcos and 140 million phones in the United States.

1976 X.25 protocol defined for public packet-switched networks.
Jobs and Wozniak start Apple Computer.

1977 Apple introduces the Apple II.
Commodore introduces the PET.
Radio Shack introduces the TRS-80.
FCC authorizes experimental wireless telephone licenses for AT&T and Motorola.

1978 Epson introduces the TX-80 dot matrix printer.
Intel invents the 8086 processor.
The 5.25 mini-floppy becomes the industry standard.

1979 Motorola 68000 is available.
Apple II Plus is introduced.
TI 99/4 is introduced.
Hayes markets its first modem.

1980 First full deployment of Signaling System 7.

1981 DOJ suit filed in 1974 against AT&T goes to trial.
The IBM PC is introduced.
Microsoft formally incorporates.
HP introduces the first 32-bit chip.
Commodore ships the VIC20.

1982 1974 AT&T suit is settled by an agreement known as the Modified Final Judgment (MFJ) and AT&T agrees to divestiture.
AT&T introduces the 5ESS switch.
Lotus Development is founded.
Compaq Computer Corp. is founded.
Commodore 64 is introduced.
Epson introduces the first notebook computer the HX-20.
TCP and Internet Protocol (IP) are established as the protocol suite for ARPAnet.

1983 GTE buys Southern Pacific's switched private network system and replaces the copper wire with fiber-optic cables.
FCC grants commercial wireless license to AT&T.
Long Distance Discount Service (LDDS) is founded.
ARPAnet is split into Military and Civilian sections, this is the beginning of the Internet.
ARPAnet switches from NCP to TCP/IP.
Apple IIe is introduced.

1984 AT&T implements divestiture plan and the Bell system ceases to exist. In its place are 22 Bell Operating Companies (BOCs), 7 Regional Bell

Operating Companies (RBOCs), and AT&T.
AT&T retains Western Electric and Bell Labs.
British Telecommunications (BT) is privatized.
3.5-inch diskette is introduced.
Stallman starts the GNU Project.
Dell Computer is founded.
Apple Macintosh is introduced.
The Tandy 1000 is introduced.
There are approximately 1000 Internet hosts.

1985 United Telecom buys 30% of GTE's fiber-optic network.
General Electric (GE) purchases RCA.
Gateway is founded.

1986 United Telecom and GTE combine their fiber networks to form US Sprint.
GE acquires NBC.
Capital Cities acquires ABC.
Apple introduces the Mac Plus.

1987 Hayes demonstrates its Integrated Services Digital Network (ISDN) adapter, a modem for ISDN lines.
The Mac SE is introduced.
Apple introduces HyperCard.
IBM introduces the PS/2.
The number of Internet hosts passes 10,000.

1988 First transatlantic fiber-optic cable linking North America and Europe is completed.
Apple files copyright infringement lawsuit against Microsoft and HP.
Western Union Telegraph is reorganized and changes name to Western Union Corp.
Approximately 45 million PCs in use in the United States.

1989 United Telecom buys the controlling interest in U.S. Sprint.
More than 100 million computers in use worldwide and the number of Internet hosts passes 100,000.

1990 The first commercial dial-up provider of Internet access, The World, comes online.
ARPAnet ceases to exist.
First World Wide Web software is created by Berners-Lee.
Gosling creates Java.
There are more than 5 million wireless subscribers.

1991 AT&T acquires NCR Corp.
Torvalds announces Linux version 0.02.
Gopher is developed at the University of Minnesota.
The World Wide Web is launched.
Western Union Corp. sells off business units and changes name to New Valley Corp.
Over 600,000 Internet hosts.

1992 United Telecom buys remaining interest in U.S. Sprint and changes name to Sprint Corporation.
MicroSoft and IBM sever the connection between the two companies.
More than 10 million wireless subscribers.
Over 1 million Internet hosts.

1993 Sprint merges with Centel Corp.
Sprint forms an alliance with Call-Net and markets services in Canada as Canada Sprint.
BT acquires a 20% share of MCI.
LDDS acquires IDB WorldCom.
Intel releases the Pentium processor.
Court rules against Apple in 1988 suit. Apple appeals.
New Valley Corp. enters bankruptcy and sells off last of its major holdings.
FCC receives authorization to auction personal communications (PCS) licenses.
Number of Internet hosts passes 2 million and there are at least 500 HTPP servers.

1994 Sprint enters an agreement with Telefonos de Mexico.
MCI forms a joint venture, Avantel, with Grupo Financiero Banamex-Accival.
AT&T acquires McCaw Cellular Communications and renames the business AT&T Wireless.
Netscape Communications founded.
Commodore Computers files bankruptcy.
Symantec and Central Point Software merge.
Appeals court rules against Apple.
FCC begins spectrum auctions with Personal Communications Services (PCS) Narrowband auction.
IBM and Motorola announce PowerPC 601 processor.
Number of Internet hosts greater than 3 million.
2847 Local service providers in the United States.

1995 Sprint forms joint venture with Tele-Communications, Comcast, and Cox Cable to buy rights to PCS licenses. Sprint enters alliances with Deutsche Telekom and France Telecom.
MCI acquires Nationwide Cellular and SHL Systemhouse.
LDDS acquires WilTel. LDDS changes name to WorldCom Inc.
AT&T reorganizes.
Sun Microsystems introduces Java.
Westinghouse buys CBS, sells manufacturing assets and takes CBS name.
Scientific Applications International Corp. (SAIC) acquires Network Solutions.

ESCOM buys Commodore's assets.
IBM buys Lotus Development.
U.S. Patent office overturns decision to grant Hyatt patent on microprocessor. Boone is recognized as prior inventor of microprocessor.
Internet Protocol, Version 6 (IPv6) is released.
Spectrum auctions continue.
3058 Local service providers in the United States.

1996 Congress passes the Telecommunications Act which effectively deregulates the United States telecommunications industry.
Southwestern Bell and Pacific Telesis merge.
NYNEX buys Bell Atlantic.
AT&T spins off Lucent Technologies.
MCI and News Corporation form a joint venture, American Sky Broadcasting. MCI forms strategic alliances with Microsoft Corp., Digital Equipment Corp., and Intel Corp.
BT purchases remaining 80% of MCI.
WorldCom buys MFS Communications.
Gannett buys Multimedia Entertainment.
Time Warner merges with Turner Broadcasting (CNN).
U.S. West acquires control of Continental Cablevision.
News Corp. buys New World Communications Group Inc.
CBS buys Infinity Broadcasting.
Disney merges with Capital Cities/ABC.
Seagate and Conner Peripherals merge.
Advanced Micro Devices and NexGen merge.
Apple buys NeXT.
Silicon Graphics and Cray Research merge.
Intel releases Pentium Pro.
IBM and Motorola ship PowerPC 603e processor.
Spectrum auctions continue.
Number of Internet hosts passes 9 million.
3832 Local service providers in the United States.

1997 BellCore is purchased by SAIC.
AT&T spins off NCR Corp.
NEC Corp and Packard Bell merge.
Spectrum auctions continue.
Number of Internet hosts passes 16 million.
3604 Local service providers in the United States.

1998 WorldCom aquires MCI. Combined company is renamed MCI WorldCom.
Compaq acquires DEC.
AT&T acquires Teleport Communications Group.
DOJ and twenty states file antitrust case against Microsoft; in October. *U.S. v. Microsoft* begins.
IEEE ratifies 802.3z as the Gigabit Ethernet standard.
Spectrum auctions continue.
There are 4144 local service providers in the United States.
94% of United States households have telephone service, 98% have a TV, and 67% cable service, 42% have a personal computer, 74% of these have a modem, and 65% have Internet access.

1999 SAIC changes Bellcore name to Telcordia Technologies.
WorldCom attempts to acquire Sprint. Acquisition is blocked by the DOJ.
AT&T merges with Tele-Communications Inc.(TCI) and TCI is renamed AT&T Broadband and Internet Services.
AT&T acquires IBM Global Network.
Aliant merges with ALLTEL.
Global Crossing Ltd. merges with Frontier Corporation.
Disney acquires Infoseek Corp. and spins off Go.com.
AOL acquires Netscape.
Testimony in Microsoft trial ends.
Spectrum auctions continue.
More than 86 million wireless subscribers.

2000 U.S. District Court orders Microsoft split in two. Microsoft appeals.
FTC approves merger of AOL and Time Warner, FCC approval required before the merger is final.
AT&T and BT form a joint venture, Concert.
AT&T announces restructuring plan to be fully implemented by 2002. After the plan is complete AT&T Business, AT&T Wireless, and AT&T Broadband will be separate businesses.
Bell Atlantic merges with GTE. GTE's Internet unit is renamed Genuity and spun off.
CBS and Viacom merge.
Judge in Microsoft case orders the breakup of Microsoft. Microsoft appeals.
Spectrum auctions continue.
Over 100 million wireless subscribers.

2001 Up to mid-February:
FCC approves merger of AOL and Time Warner.
Deutsche Telekom AG announces intent to sell its 10% stake in Sprint.
Lucent Technologies announces plans to sell its fiber-optic unit.
SWIFT announces it has entered into an agreement to subcontract its network operations to Global Crossing.
Disposable cellular phone expected to be introduced.

Microsoft case: oral arguments begin before the U.S. Court of Appeals.
Spectrum auctions continue.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SATELLITE SYSTEMS • COMMUNICATION SYSTEMS, CIVILIAN • COMPUTER NETWORKS • DATA TRANSMISSION MEDIA • DIGITAL SPEECH PROCESSING • OPTICAL FIBER COMMUNICATIONS • RADIO SPECTRUM UTILIZATION • TELEPHONE SIGNALING SYSTEMS, TOUCH-TONE • VOICEBAND DATA COMMUNICATIONS • WIRELESS COMMUNICATIONS • WWW (WORLD-WIDE WEB)

BIBLIOGRAPHY

Brock, G. W. (1998). "Telecommunication Policy for the Information Age," Harvard University Press, Cambridge, Massachusetts.
Casson, H. N. (1910). "The History of the Telephone," A. C. McClurg, Chicago.
Keagy, S. (2000). "Integrating Voice and Data Networks," Cisco Press, Indianapolis.
Kurose, J. F., and Ross, K. W. (2001). "Computer Networking: A Top-Down Approach Featuring the Internet," Addison-Wesley, Reading, Massachusetts.
Lin, Y.-B., and Chlamtac, I. (2001). "Wireless and Mobile Network Architectures," John Wiley & Sons, New York.
Pildush, G. D. (2000). "Cisco ATM Solutions: Master ATM Implementation of Cisco Networks," Cisco Press, Indianapolis.
Schulzrinne, H., guest ed. (1999). "IEEE Network," Special Issue on Internet Telephony, Vol. 13, No. 3.
Valovic, T. (1992). "Corporate Networks: The Strategic Use of Telecommunications," Artech House, Boston.



Telephone Signaling Systems, Touch-Tone

A. Michael Noll
University of Southern California

Leo Schenker
AT&T Bell Laboratories

I. Introduction
II. History of Development and Introduction
III. Touch-Tone Dialing Scheme
IV. Touch-Tone Dial Telephone
V. Touch-Tone Receiver
VI. Conclusions

GLOSSARY

Common control Central-office switch type of system indicating that switching connections are set up by common equipment once the complete dial information has been received.
Guard action Means for reducing the probability of talk-off.
Loop That part of the telephone plant that connects subscribers' premises to the local central office.
Multifrequency key pulsing (MFKP) Method used for signaling between telephone switches.
Progressive control *See* Step-by-step.
Station set Telephone set.
Step-by-step Central-office switching equipment of the type where the incoming dial pulses directly affect the position of switches and thus set up a talking path.
Talk-off Imitation of control signals by speech.

TOUCH-TONE dialing is a voice frequency means of signaling from telephone sets to central-office switches to indicate the destination of a call. The user interface is a set of pushbuttons. Compared to the older dialing method with the rotary dial and dial pulses, Touch-Tone signals have frequency and level characteristics such that they can be transmitted beyond central offices to any point in a telephone network. Hence, Touch-Tone signals can be used for the transmission of data after a connection has been set up. Touch-Tone dialing is widely used in North America and many other countries.

I. INTRODUCTION

Touch-Tone dialing is a method of sending signals from telephone customers' premises to central offices and beyond. It was first introduced in 1964. Compared to rotary

dialing of dial pulses, its principal advantages are as follows.

1. All the signaling energy is in the voice frequency band, making it possible to transmit signaling information to any point in the telephone network to which voice can be transmitted, that is, "end to end."
2. Touch-Tone dialing is faster, reducing the dialing time for users and, equally important, reducing the holding time for central-office common equipment.
3. It provides a means for transmitting more than 10 distinct signals: 12 in all standard implementations (and the scheme is capable of 16).
4. It is a more convenient signaling method, as attested by users.

II. HISTORY OF DEVELOPMENT AND INTRODUCTION

The developmental work that led to the introduction of Touch-Tone dialing began at Bell Laboratories in the mid-1950s. This was not the first occasion on which the merits of a "push-button" dial (as it was then called) were studied. Following the introduction of operator multifrequency key pulsing (MFKP), a system of ac signaling for station equipment was designed, built, and tested in the laboratory in 1941. World War II interrupted further work, and 7 years elapsed before equipment was installed and used on an experimental basis in a small trial at Media, Pennsylvania, in 1948. This system was based on the transmission of damped oscillatory signals at two of an available six frequencies and required special receivers at the central office to detect and convert the signals. The dialing unit had mechanical linkages that plucked two of six metal reeds, each of which was resonant at a specified frequency. When a customer pushed any 1 of the 10 buttons, two reeds were plucked to form a signal coded to the corresponding digit. The energy so generated was transferred inductively to coils in the station-set network and so transmitted to the receiver at the central office. Although this mechanism was cumbersome, the performance of the equipment and the reaction of the customers pointed the way to an ultimately feasible system and indicated a favorable public response to push-button signaling. The technical and economic aspects of the system tested at Media were not attractive, and further work was deferred.

The year of the Media trial, 1948, is also noteworthy because in June of that year the discovery of the transistor was announced. Concepts that formerly were considered exotic became practical. The activity in electronic device development resulted in an entirely new array of circuit components. Miniature capacitors, ferrite inductors, precision resistors, and printed wiring became available, and electric power consumption ceased to be a formidable obstacle.

Using this new technology, a compact multifrequency oscillator, equipped with push buttons for selecting and controlling voice frequency signals, was developed in the second half of the 1950s. The oscillator was particularly adapted to the low and variable power available from the central-office battery over the range of existing loops to the station set. The oscillator design provided for control of the signal at a level high enough to be reliably detectable at the central office in the presence of the noise from numerous sources that is always present on telephone circuits, but not so high as to exceed crosstalk requirements. The concept of a four-by-four frequency code resulted in a relatively simple mechanical system at the station set.

Concurrently with these electrical and mechanical developments, human factors engineers studied pertinent psychophysical factors and performance ratings of button arrangement were made. The optimum size, spacing, travel, and operating force of the buttons were determined. It was also established that feedback of the signal tones through the telephone receiver was desirable. These studies indicated that customer dialing could be speeded up and generally facilitated by the use of a push-button operation, probably without increasing dialing irregularities seriously.

On the basis of satisfactory results in the areas of electromechanical design and psychophysical studies, a moderate-sized field trial of push-button dialing was designed and carried out in 1959. Two more trials were to follow before the system was introduced for general use. The chief objective of the first trial was to determine the effectiveness and ease of use of a modern button-operated mechanism with then-current types of switching systems, when placed in the hands of a typical segment of the public. Two locations were selected. One of these, at Elgin, Illinois, was equipped with fast-operating common control equipment; the other, at Hamden, Connecticut, used the slower step-by-step switching equipment. The necessary central-office receiver and converter equipment was provided on a "black-box" basis. At Hamden, this equipment included a digit storage facility and a dc outpulser geared to the speed of the step-by-step system. Approximately 120 customers in each central-office area were provided with Touch-Tone sets of a preliminary design.

As in the Media trial, customers in both areas were enthusiastic about this experimental service, basing their reaction on the increased speed and general ease of use. In a relatively short period, the customers achieved a dialing speed with the Touch-Tone sets that was almost twice their

established rotary-dial speed. It was anticipated from laboratory tests that customer signaling irregularities would tend to occur at a higher rate with push buttons than with the slower and more familiar rotary dial, and this was confirmed in the field trial. Within a relatively short period, the Touch-Tone error rate of the trial customers showed a trend that led to the projection that, with longer experience, accuracy equal to that of rotary dialing would be achieved.

Concurrent development work on the electromechanical design of Touch-Tone dials and receiver-converter equipment resulted in designs proposed for manufacture, and a second set of trials, technical trials of this equipment, was planned and carried out in 1960. As in the Hamden and Elgin trials, two locations were selected. One of these was at Hagerstown, Maryland, served by common control switching equipment; the other was the Cave Spring office at Roanoke, Virginia, served by step-by-step equipment.

The results of this second set of trials led to the conclusion that a practical new dialing scheme could be developed based on the technical concepts of the system and components used in the technical trials. The technical performance of station sets and central-office equipment brought no surprises. Signal imitations from speech (a problem known as talk-off) and other noises were not a problem. Customers continued to be highly supportive.

Touch-Tone calling was introduced as a premium service on a market-trial basis at Findlay, Ohio, and at Greensburg, Pennsylvania, in 1961. The equipment supplied for the station installations in these areas included some minor refinements relative to that used in the technical trials, and some modifications of the dial were made to adapt the design to the full line of station sets, all of which were equipped with Touch-Tone dials. Figure 1 is a photograph of the Touch-Tone desk set used in the market trials. The appearance design was by Henry Dreyfuss. The central-office equipment also included minor refinements, but was essentially unchanged.

Central offices beyond those in the market trial were converted for Touch-Tone dialing beginning in 1963. The new method of dialing was available at the New York Worlds Fair in 1964. Today, all central offices accommodate Touch-Tone dialing.



FIGURE 1 Early Touch-Tone desk telephone. [Courtesy of AT&T Bell Laboratories.]

III. TOUCH-TONE DIALING SCHEME

In the previous section, several advantages of Touch-Tone dialing over rotary dialing were listed. When the development program that led to Touch-Tone dialing began in 1953, not all of these were considered equally important. Initially, the only objective was to find a way to reduce dialing time, probably through a scheme that would involve a push-button person–machine interface. It was known from the Media trial that customers would be favorably disposed toward push-button dialing.

The development environment was governed by the following factors. The rotary dial had been around for several decades and was very inexpensive. The transistor had been invented a few years earlier; no mass applications of this new device existed. Power dissipation, variation of gain among devices, reliability, and cost were all issues. Thus, it was risky to consider a scheme that would involve an oscillator in millions of telephones and in millions of diverse locations. Some signaling method not requiring an active device was likely to be more practical. For example, one scheme that was studied involved the generation of damped oscillatory waves by interrupting the direct current through the coil of an inductor-capacitor tuned circuit. By switching the capacitor to 1 of 10 taps on the inductors, 10 different digit-identifying frequencies could be generated, all in the telephone voice frequency range. But such signals were highly likely to be imitated by the user's (inadvertent) speech in the dialing interval. Hence, in this scheme, each push-button operation also caused a stepwise reduction in the direct current (dc) drawn by the telephone from the central office. The damped oscillations were ignored by the central-office receiver unless accompanied by a dc step.

Very soon it became clear that it was essential to be able to transmit customers' signals "end to end," that is, from one customer's premises, through one or more central offices, to the premises of another customer, anywhere that voice can be sent. Two requirements resulted from the end-to-end signaling objective.

1. The signals must not contain an out-of-band component such as a dc step.

2. Sustained rather than damped signals must be used to maintain adequate signal-to-noise margins for the wider range of transmission losses when two customer loops are involved.

The first of these two requirements—the need for the signals to be wholly contained within the voice frequency band—also brings with it the problem of vulnerability to talk-off. The second reintroduced the uncharted domain of active devices.

A. Choice of Code

When only voice frequencies are employed, protection against talk-off must rely heavily on statistical tools. This protection is required only during interdigital intervals; speech interference with valid signals can be avoided by transmitter disablement when a push button is operated. Since signals with a simple structure are prone to frequent imitation by speech and music, some form of multifrequency code particularly difficult to imitate is indicated. If the signal frequencies are restricted in a binary fashion to being either present or absent, the greatest economy in frequency space results from the use of all combinations of N frequencies, yielding $n = 2^N$ different signals. However, some of these are no more than single frequencies and are therefore undesirable from the standpoint of talk-off. Another drawback is that as many as N frequencies must be transmitted simultaneously; these involve an N-fold sharing of a restricted amplitude range and may also be costly to generate. If $n = 10$, N needs to be at least 4. At the expense of using up more frequency range, one is led to a P-of-N code, yielding $n = N!/P!(N - P)!$ combinations. There is a statistical advantage in knowing that there are always P components in all valid signals, no more and no less.

To minimize the number of circuit elements, as well as to reduce the sharing of amplitude range, P should be as small as possible, yet larger than unity for the sake of talk-off protection. Let us, then, examine codes in which $P = 2$. If one can be found that is not readily imitated by speech or music, there is no merit in choosing P higher than 2, provided that the total number of frequencies N needed for the required number of combinations can be accommodated in the available frequency spectrum. With $P = 2$, N must be at least 5 to provide 10 combinations (for the 10 digits). With $N = 6$, 15 combinations are available. The MFKP signaling scheme makes use of a two-of-six code. There were reasons to assume that MFKP would not have adequate talk-off performance. As discussed below, a code was chosen with attributes that provided protection against talk-off by several means and with which less than 1 talk-off in 5000 calls was achieved, a degree of talk-off performance that was deemed minimal for customer acceptance.

There are advantages in imposing the further restriction that, with $P = 2$, the frequencies for each combination fall, respectively, into two mutually exclusive frequency bands. If, for example, 15 or more combinations are required, N must be at least 8. In the four-by-four code, eight signal frequencies are divided into two groups: group A, the lower four frequencies; and group B, the upper four. Each signal is composed of one frequency from group A and one from group B, resulting in 16 signal combinations.

With a two-group arrangement, it is possible at the receiver to separate the two frequencies of a valid signal by band filtering before attempting to determine the two specific components. This separation of the two components of a signal renders reliable discrimination between valid signals and speech or noise.

B. Choice of Frequencies

Attenuation and delay distortion characteristics of typical combinations of transmission circuits were such that it is desirable to keep the frequencies of a telephone signaling system within the 700- to 1700-Hz range.

The choice of frequency spacing depends in part on the accuracies of the signal frequencies. It was expected that signals generated at the station set could be held within 1.5% of their nominal frequency values and that the pass bands of the receiver selective circuits could be maintained within 0.5% of their nominal ranges. On the basis of these numbers, the selective circuits of central-office receivers need to have recognition bands of at least ±2% about the nominal frequencies.

The standardization of amplitude at the output of the limiters used in the receiver at the central office (Fig. 2)

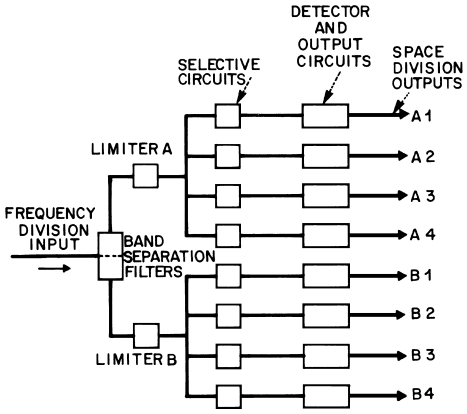


FIGURE 2 Simplified diagram of the book Touch-Tone receiver used in central-office switching systems.

permits an accurate definition of recognition bands in the receiver, independently of levels at the receiver input. As a result, frequencies may be spaced closely, approaching the recognition bandwidth of 4%. If the lowest frequency is chosen as 700 Hz, the next frequency must then be more than 728 Hz. Wider spacing makes the precise maintenance of the bandwidth less critical.

Another factor can profitably be taken into account in the selection of a frequency spacing. To reduce the probability of talk-off, the combinations of frequencies representing bona fide signals should be such that they are not readily imitated by the output from the speech transmitter. In a receiver with the guard action described, no sound composed of a multiplicity of frequencies at comparable levels is likely to produce talk-off. Thus, consonants present no problem. Vowels, however, do present a problem, as do single-frequency sounds such as whistles that are loud enough to encounter some harmonic distortion in a carbon transmitter. It was shown by H. Fletcher that an electrical analogue of the mechanism involved in the articulation of vowels is a buzzer (the vocal chords producing a fundamental and a long series of harmonics) followed by a selective network that shapes the harmonics into formants of the vowel sound. As a result, the spectrum of any sustained vowel contains a number of frequencies bearing harmonic relationships to each other. Hence, it is desirable that the pairs of frequencies representing valid signals avoid as many of these harmonic relationships as possible.

A family of frequencies that avoids a large proportion of troublesome combinations and also meets all the other requirements discussed so far is as follows, the adjacent frequencies in each group being at a fixed ratio of 21:19, with 2.5 times this interval between the groups.

Group A (Hz)	Group B (Hz)
697	1209
770	1336
852	1477
941	1633

All frequencies are essentially within the 700- to 1700-Hz range, and the spacing is adequate to accommodate the recognition bands.

The 16 pairs of frequencies representing valid signals avoid low-order ratios. This is illustrated in Fig. 3, where frequencies are plotted on two logarithmic scales: those below 1000 Hz as ordinates and those above 1000 Hz as abscissae. Any valid signal is represented by a pair of coordinates, namely, its two component frequencies. The 16 square windows represent the ±2% recognition bands required.

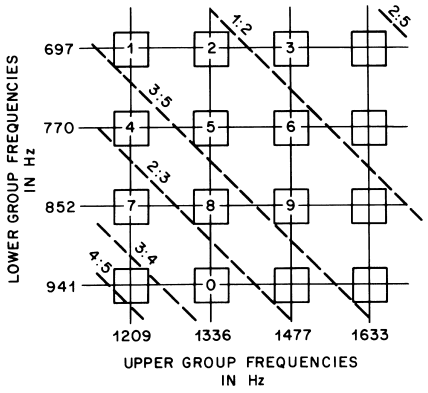


FIGURE 3 Window diagram for signal frequencies.

The diagonal lines in Fig. 3 are the loci of pairs of frequencies having simple harmonic relationships, that is, 1:2, 2:3, etc. The avoidance of these particular diagonals is beneficial because they represent the effects of harmonics not only of the corresponding order but also of higher order, for example, the third and sixth, the ninth and fifteenth, and so on. Not all applicable diagonals are shown in Fig. 3, and 18 of the 65 potentially troublesome combinations are not avoided. However, if two frequencies in speech representing two fairly high-order harmonics (e.g., 4:7 and 7:12) happen to fall at proper signal frequencies, their fundamental will be low in frequency. The odds are then good that other harmonics within the telephone speech band will together have sufficient relative intensity to bring guard action into play.

Comparable benefits with respect to talk-off can be derived from other sets of frequencies with the same geometric spacing but displaced up or down on the frequency scale. The effect of such displacement is to shift all the windows in a direction parallel to the diagonals.

C. Choice of Amplitudes

Since signaling information does not bear the redundancy of spoken words and sentences, yet must be transmitted with a high degree of reliability, it is advantageous for the signal power to be as high as permitted by the environment. A nominal combined signal power for the two frequencies of 1 dB above 1 mW at the telephone-set terminals was adopted as a realistic value, a ±3-dB tolerance about this nominal value being acceptable.

For subscriber loops, the maximum slope between 697 and 1633 Hz is about 4 dB, the attenuation increasing with the frequency. In two subscriber loops (as in end-to-end signaling), the slope may be 8 dB, although, statistically, this is unlikely. A reduction in the maximum level difference at the receiver in the two signal components

can be achieved by transmitting the group B frequencies at a level 3 dB higher than that of the group A frequencies. In this way, the nominal amplitude difference at the receiver input between the two components of a valid signal is never more than 5 dB end to end and 1 dB at the central office. The nominal output powers were chosen as -3.5 and -0.5 dBm for groups A and B, respectively, adding up to 1.3 dBm. In more than 99% of station-to-central office connections, the 1000-Hz loop loss was expected to be less than 10 dB (during the early 1960s). Similarly, the station-to-station loss at 1000 Hz was estimated to be less than 27 dB in more than 99% of all connections. Making some allowance for slope and variations in the generated power, the minimum signal power at a central-office receiver was estimated to be 15.5 dBm. At a receiver involved in end-to-end signaling, the minimum power was estimated to be 32.5 dBm. Higher minimum levels can probably be expected today.

D. Signal Duration

In the development of the Touch-Tone dialing system, two important parameters had to be fixed before adequate data were available: the duration of a valid signal and the signaling rate. The probability of talk-off can be reduced by increasing the duration of the test applied to a signal by the receiver before accepting the signal as valid. It was clearly undesirable, however, to require even the most adept users to extend a push-button operation beyond an interval determined by their innate or acquired mental and physical acuity. A frame of reference for setting requirements for these intervals can be formed in two ways. First, competent typists achieve a speed of 60 words per minute, and, at five letters per word, the corresponding time interval between letters is 200 msec. Second, musical compositions occasionally include 64th notes; although the speed of performance is at the discretion of the performer or conductor, the duration of a 64th note on a piano is estimated to be rarely as short as 45 msec. A minimum of 40 msec was set as the objective for both signal and intersignal intervals, allowing for a dialing rate of more than 10 signals per second.

The allocation of a minimum of 40 msec to these intervals proved to be reasonably sound. The median tone time found in the field trials was approximately 160 msec; durations of 30 to 40 msec were observed, but only approximately 0.15% of dialing starts failed because of short tones. The median interdigital time was approximately 350 msec; no case of failure because of short interdigital intervals was observed. Fortunately, consideration never had to be given to increasing the minimum signal time to improve talk-off performance.

IV. TOUCH-TONE DIAL TELEPHONE

The scheme described in the previous section would have remained just another scheme if it had not been feasible to develop a dial that could be manufactured and maintained at a reasonable cost. The Touch-Tone dial represented a bold departure from past practices in a product line that today would be called consumer products. A proven mechanical device, honed to near-perfection as a result of extensive use and several redesigns, was to be replaced by a new mechanism and circuitry, including a relatively immature electronic active component, the germanium transistor, to be installed in millions of homes and offices.

A. Human Factors Design

Extensive tests were performed to determine customer preference, dialing speed, and error rate as various physical design parameters were changed. These parameters included button size and spacing, stroke, force displacement characteristics, and many others. One important parameter, although included in the test programs, ended up being determined by mechanical design considerations: the button arrangement. A rectangular array of three rows of three buttons, plus one button in the center of the fourth row ($3 \times 3 + 1$), was selected because it resulted in an arrangement that required fewer electrical contacts. Of course, later, with the addition of the $*$ and # buttons, this became a 3×4 array. Fortunately, this button arrangement turned out to be close to optimal, in both performance and preference.

Consideration was given to numbering the button in the same sequence as an adding machine, that is, 7, 8, and 9 in the top row, the "Sundstrand" arrangement. User preference was the reverse order of row numbers, that is, 1, 2, and 3 in the top row. Not enough people were users of adding machines in the late fifties; handheld calculators did not exist.

User preference and performance studies, coupled with design considerations, eventually resulted in the now-familiar attributes of desk-type Touch-Tone telephones: $\frac{3}{8}$-in. square buttons, separated by $\frac{1}{4}$ in.; a $\frac{1}{8}$-in. stroke without snap action; and about $100g$ force at the bottom of the stroke.

B. Initial Mechanical Design

The push-button mechanism was the means by which the customer controlled the operation of the dual-frequency oscillator. The initial design involved a complex, yet ingenious, combination of levers and switches. In the early part of the downstroke of a push button, each tuning capacitor

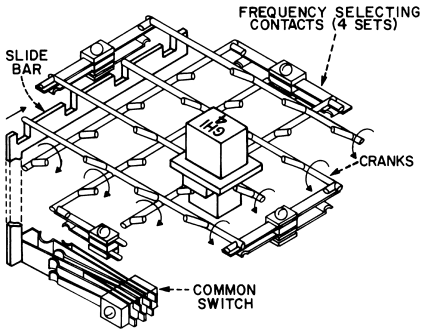


FIGURE 4 Mechanism schematic for the initial Touch-Tone caller.

was connected to a tap on its associated tuning coil to select the two frequencies of the signal for that button. Later in the downstroke, a switch common to all buttons was operated. The common switch had the following four functions.

1. Attenuation was inserted into the receiver circuit to reduce the level of side tone during signaling to a comfortable value. This had to occur first.
2. The transmitter path was opened to prevent speech or background noise from interfering with the signal.
3. Power was applied to the transistor oscillator.
4. The dc through the tuning coils was interrupted to initiate the signal at full amplitude by shock excitation. This had to occur last so that none of the energy stored in the tuning coils was wasted prior to enabling the transistor.

In the initial design a mechanism was conceived that kept the number of frequency selection contacts to a minimum (7), rather than to have each button operate two frequency selection contacts, which could have required 20 and, later, 24 contacts. This type of mechanism was used in the majority of Touch-Tone sets.

A schematic of the mechanism is shown in Fig. 4. Two sets of cranks operated the coil-tap switches. The cranks in the column direction operated switches connected to the high-band coil, and the cranks in the row direction operated the low-band switches. When a button was pushed, the pair of cranks crossing below that button was rotated, and one coil tap contact was made for each tuned circuit, thus selecting the two appropriate frequencies. An arm on each of the "row" cranks moved a slide bar when any one of them was rotated. Movement of this slide bar operated the common switch. Figure 5 shows the front and back of an early Touch-Tone dial.

C. Evolution of the Touch-Tone Dial

The circuitry and mechanism of the Touch-Tone dial have evolved significantly from the early versions. The availability of silicon and tantalum thin-film integrated circuits

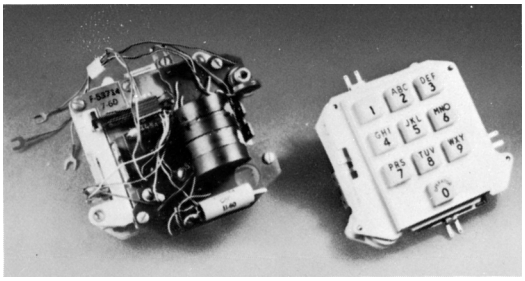


FIGURE 5 Initial Touch-Tone dial assembly. [Courtesy of AT&T Bell Laboratories.]

led to a second generation of designs. Physically, this consisted of two glass substrates glued together, with about 20 leads to the crank mechanism and the speech network. Most transistors, most diodes, and some resistors were implemented in a silicon integrated circuit chip, with another smaller chip for amplitude-regulating diodes. The tuning elements were capacitors and resistors implemented in tantalum thin-film form. As technology evolved, it became possible to eliminate the separate diode chip and to combine the two glass substrates into a single ceramic substrate.

The third and current generation is based on very large-scale integration silicon technology—hundreds of devices on a chip—and is produced by a variety of manufacturers. In some cases, the very large-scale integration chips provide functions beyond the generation of the Touch-Tone signals, for example, storage of telephone numbers that can then be transmitted by pushing one or two buttons. Some of the chip designs can also generate the rotary dial-type dc pulses.

The mechanical design of Touch-Tone dials has also changed. There are no more crank-type actuating mechanisms. Rubber dome keypads or membrane keypads are used in the latest designs. The frequency selection and common switch functions are effected by the closing of a single contact under a push button and the use of logic circuitry.

V. TOUCH-TONE RECEIVER

Prior to the introduction of automatic switching, the means of "signaling" the destination of a call involved the spoken language or, more accurately, languages. In areas where more than one language was in widespread use, multilingual operators were required at switchboards. The advent of dialing simplified this situation because dialing became a universal people machine "signaling language." With the introduction of a second customer signaling scheme, the earlier multilingual requirement for operators was now imposed on the central-office signal reception equipment.

Historically, automatic switching systems can be divided into progressive control and common control systems. In progressive (step-by-step) systems, the dc pulses resulting from the operation of a rotary dial directly control a series of switches, one for each digit of the dialing sequence, to establish the talking path between calling and called customers. In common control systems the information transmitted by the dialing signals is stored in a register and action to establish a talking path is taken only after the information for the complete dialing sequence has been received. Progressive switching systems are obsolete and have been replaced by modern electronic systems, all of which use common control and registers.

A. Receiver Design

Figure 6 is a schematic diagram of the Touch-Tone receiver. A Touch-Tone signal first passes through a high-input impedance amplifier. This amplifier also includes a high-pass filter to attenuate the 60- and 180-Hz components of induced power line noise and to attenuate the dial tone, standardized as a combination of 350 and 440 Hz.

After amplification, the two signaling frequencies are separated into their respective groups by the band elimination filters.

The response of each limiter to the single Touch-Tone frequency at the band elimination filter output is a symmetrical square wave of fixed amplitude containing the fundamental and odd harmonics of the signal frequency. Each limiter drives a group of circuits used for signal frequency recognition. In each group, the tuned circuit corresponding to the incoming signaling frequency responds to the fundamental of the square wave with an amplitude sufficient to operate a detector. However, the receiver does not deliver output signals the instant the detectors operate. Instead, it makes a check to determine the presence of one and only one valid tone in each of the two-frequency groups. This check must continue uninterrupted for a prescribed time interval before an output signal duration test is performed by the checking circuit and signal timer that are shown in the block diagram in Fig. 6.

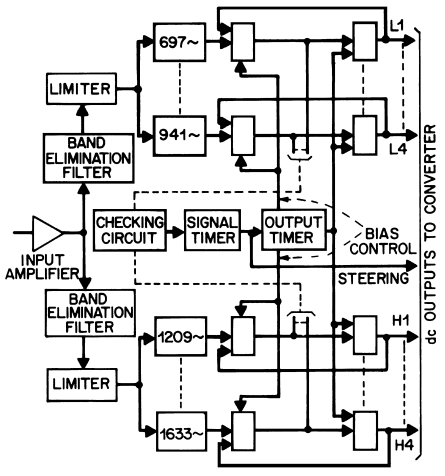


FIGURE 6 Schematic diagram of the Touch-Tone receiver.

The check for one and only one frequency in each group is achieved by a combination of simple OR and AND logic circuitry and a characteristic of the limiter permitting the operation of only one detector in a group at a time. When two frequencies of comparable amplitude are applied to a limiter input, neither frequency has sufficient energy at the limiter output to operate a detector.

Once the signal duration test is completed, the receiver is ready to deliver output signals. On minimum-length input signals, there is very little time left after tuned circuit buildup and the signal duration test for delivery of output signals. Clearly, some sort of memory is required to prolong the output signals. In the Touch-Tone receiver, this memory is achieved by providing an output timer and output gates with positive feedback to the associated detector circuits.

In addition to the channel outputs (one per signal frequency), there is a steering output that controls the advance of steering logic in the attached digit registering equipment from one digit storage position to the next. The steering lead remains energized as long as output signals are present or signal tones are present, whichever is longer. On short signals, the locked-in detectors hold the checking circuit and signal timer on so that the steering lead remains energized until after the output timer goes off. On long input signals, when the output timer goes off, the detectors remain operated from the tuned circuit outputs and hold the steering output in the operated condition until the end of the Touch-Tone pulse. This avoids double registration of long signals.

The recognition bandwidth of each channel circuit is dependent on the detector threshold level, the amplitude of the fundamental component of the limiter output, and the Q of the tuned circuit. A recognition bandwidth of approximately 5% of the nominal channel frequency is required to allow for variations in received frequencies and receiver frequency response. Knowing the required recognition bandwidth, the most important of those variables to be determined is the Q of the tuned circuit.

Receivers designed according to these principles were used in the various trials, and performance was good, including digit simulation performance, which was 1 talk-off in 10,000 calls. (The rate of digit simulation was poorer when the simulation of the six unused signals is also taken into account.)

B. Evolution of the Touch-Tone Receiver

The architecture of the receiver has essentially remained unchanged over the years. However, the physical design has changed radically, as would be expected in light of the technological advances since the 1960s. The receiver used in the various trials and in production through about 1975 used all discrete components, 50–60 transistors, coils, and so on. The very early receivers took up about 6 in. of height on 23-in.-wide racks. Soon a redesign resulted in two central-office receivers being accommodated in the same space.

During the initial phase, a receiver intended for use in PBXs was developed with a simplified architecture to save space and cost. This was justified on the basis that amplitude sensitivity and talk-off requirements could be less stringent. However, in many PBX applications this turned out to be unsatisfactory and the receivers were replaced by central-office receivers with special mounting arrangements.

Receivers introduced in 1975 had the passive LC filters replaced by active RC filters. Integrated circuits on thin-film substrates incorporating resistors were used, as well as operational amplifier chips and chip capacitors. Typical physical designs involved three 5×7-in. printed circuit boards.

Receivers in the most current applications are essentially on one silicon chip. For example, for a digital central office, where advantage is taken of the fact that the analog signals offered to the receiver have been encoded into pulse code modulation format by common input circuitry, a Touch-Tone receiver is accommodated on a 40-pin digital signal processor chip that has 45,000 transistors. The program contained in the chip emulates all the signal processing functions described earlier.

VI. CONCLUSIONS

Touch-Tone dialing, conceived in the 1950s and introduced in the 1960s in the United States, has been widely accepted by the public and is the major way most telephones signal with local switching machines.

In some countries, push-button dial phones are used that emulate the dc signals produced by rotary dials. Such instruments are usually capable of storing the information input by the actuation of the push buttons and transmitting the dc pulses at the rate at which the central office can accept them. In this way, non-tone-generating push-button schemes satisfy the needs of users for more rapid and convenient dialing, but they do not significantly reduce the interval between the start of dialing and the setting-up of a connection. In fact, they may increase the interval between the end of dialing and the setting-up of a connection. The availability of inexpensive non-tone-generating push-button dials may slow down the use of Touch-Tone when Touch-Tone-capable central offices are eventually introduced in these countries.

A major advantage of Touch-Tone is the potential for data transmission and remote control. Other services requiring interaction between the telephone user and service providers, such as banks and investment companies, are also important applications of Touch-Tone.

SEE ALSO THE FOLLOWING ARTICLES

DATA TRANSMISSION MEDIA • DIGITAL SPEECH PROCESSING • NETWORKS FOR DATA COMMUNICATION • TELECOMMUNICATIONS • VOICEBAND DATA COMMUNICATIONS

BIBLIOGRAPHY

Battista, N., Morrison, C. G., and Nash, D. H. (1963). "Signaling system and receiver for Touch-Tone calling," *IEEE Trans. Commun. Electron*. **82**.

Benson, L., Crutchfield, F. L., and Hopkins, H. F. (1963). "Application of Touch-Tone calling in the Bell system," *IEEE Trans. Commun. Electron*. **82**.

Ham, J. H., and West, F. (1963). "A Touch-Tone caller for station sets," *IEEE Trans. Commun. Electron*. **82**.

Noll, A. M. (1998). "Introduction to Telephones and Telephone Systems," 3rd ed., Artech House, Boston.

Schenker, L. (1960). "Pushbutton calling with a two-group voice frequency code," *Bell Syst. Tech. J.* **39**.



Voiceband Data Communications

Whitham D. Reeve
Reeve Engineers

I. Applications and Networks
II. Voiceband Channels
III. Modems
IV. Baseband Signaling
V. Channel Equalization
VI. Passband Data Communications
VII. Future of Data Communications

GLOSSARY

Analog Time continuous signal.
Asynchronous Signaling bursts not synchronized with one another or a reference clock.
Baseband Signals in their unmodulated form, usually including spectral components near-zero frequency (dc).
Baud Signaling rate of a data signal in symbols per second.
Bit rate Information rate of a data signal, not necessarily equal to baud, measured in bits per second.
Digital Signaling with elements from a finite set.
Equalization Correction of channel transmission characteristics.
Full-duplex Simultaneous two-way communication.
Half-duplex One direction at a time communication.
Intersymbol interference Interference arising in pulse detection from the tails of neighboring pulses.
Modem MOdulator–DEModulator, device for communicating digital data over an analog channel.
Nyquist pulse Signaling pulse that requires minimal bandwidth for a specified spacing between noninterfering pulses.
Packet communication Data communication by separate packets of data individually conveyed through a network.
Passband Transmission band excluding frequencies near zero that carries a modulated communication waveform.
Private line Point-to-point communication circuit leased from a communications carrier for the exclusive use of the leasing party; also called *special access* circuit or dedicated circuit.
Pulse Elementary signaling waveform for digital communication for which the level or other parameter is selected from a discrete set.
Shell mapping Mapping data bits to signal points in a multidimensional signal constellation by partitioning the original two-dimensional signal constellation into rings containing an equal number of points.
Symbol A signal selected from a discrete set that occurs in one elementary signaling interval; a symbol may contain one or more bits.

Synchronous Signaling at uniform clocked intervals.
Transversal filter Linear filter consisting of a tapped delay line with the output composed of the weighted sum of the voltages at the taps.
Trellis coding Convolutional coding that correlates a sequence of signal subsets in a partitioned signal constellation to improve bit error rate performance, as in trellis coded modulation.
Voiceband The voice frequency band between approximately 300 and 3400 Hz.

ELECTRICAL DATA communications has a long history starting with the Morse code telegraph systems built along the railroads in the mid-19th century.[1] The telegraph was a form of digital baseband signaling using metallic media (open wire) but the rate of information transfer was low by today's standards. The transmission of speech over twisted pairs of wire followed. The bandwidth of the speech signals was restricted to less than 4 kHz due to technological limitations, but this bandwidth has carried through to the present. Studies show telephone subscribers have a 90% satisfaction level with a bandwidth of around 3.1 kHz. As telecommunication systems evolved, data were carried on purely digital facilities from one location to another. Now, practically all backbone transmission systems use digital transmission based on optical fiber. However, voiceband data communication, which is the transmission of a digital data stream through a channel designed for a single analog voice channel, still is used in the access network. The access network is that part of the network between the subscriber's premises and the serving central office. In particular, the dialup voiceband modem (MOdulator–DEModulator) is a common access device primarily because of the comparatively low cost of the modem in combination with the switched voiceband channel. Only now are these voiceband modems being supplemented and supplanted in the access network by digital subscriber line (xDSL) technology because of user's digital speed requirements and a corresponding reduction in cost of the digital technology. The xDSL, unlike the dial-up modem, has a full-time connection to the data network (primarily the internet).

Circuit-switched and dedicated analog telephone channels are common voiceband channels. Modems, which translate digital data streams into analog signals appropriate for the voiceband channel and attempt to correct the channel deficiencies, are the key component of modem data communication. Sophisticated modulation and adaptive signal processing techniques, together with very large-scale integration (VLSI) in semiconductor chips, have made it possible to transmit data through individual voiceband telecommunication channels at rates as high as 56 kb/s using the basic principles of bandwidth-efficient modulation and adaptive compensation of channel distortions.[2] These same basic principles are used in digital subscriber lines, such as xDSL, digital microwave relay systems, wireless access technologies, such as digital cellular and personal communication services (PCS), and optical fiber communications.

VLSI has greatly reduced the cost of the digital signal processing (DSP) required for relatively high-speed voiceband data transmission. Starting in the early 1980s, VLSI made it possible to inexpensively implement the filtering, equalizing, echo canceling (equalizing and echo canceling actually are filtering functions), and analog-to-digital/digital-to-analog conversion requirements of bidirectional voiceband data transmission. In addition to these integrated circuit developments there have been advancements in DSP techniques. Trellis coded modulation (TCM) is a technique used in all high-speed voiceband modems. TCM allows a higher bit rate by redundantly encoding information bits from the source, which reduces the error rate in distorted channels. The performance gains of TCM come at the expense of great computational complexity possible only with VLSI. TCM is discussed in Section VI.

I. APPLICATIONS AND NETWORKS

Data communication has long been with us, represented by written languages, semaphore signaling, and many other formats. Electrical data communication, including the concepts of multiplexing and efficient coding, was developed in the telegraph systems of the 19th century through the pioneering efforts of Samuel F. B. Morse. In the first half of the 20th century, teletype and telephoto services were implemented and became essential to the publishing industry. User-to-user data communication systems such as telex also became available as important commercial services. Electrical data communication became important to society in general with the beginning of the computer era in the 1960s.

In an information society, data are what carry information among people and machines. Word processing and document facsimile transmission, airline reservations and credit card authorization systems, automatic teller

[1]This article is an update of an original article by S. Weinstein in the 1987 edition of this encyclopedia. The author of the current article relied heavily on this earlier work.

[2]The 56 kb/s rate is achievable in principle but, due to power limitations imposed by the Federal Communications Commission (FCC) in the access network to reduce the chance of crosstalk interference, the highest speeds possible are 50–53 kb/s (but actual speeds generally are lower).

machines, personal computers and, of course, the world wide web all represent and process information as data. In some cases, these data are produced and used in one isolated location, but usually data are shuttled from one machine to another through a variety of communication facilities. If these activities take place within one building, as it might in a company office, a factory, or a local computing environment, a local area network (LAN) is likely to connect the equipment together. The interconnection of LANs by metropolitan area networks (MANs) and wide area networks (WANs) and the very rapid development of the internet show the important role of data communication in removing geographic constraints on the dissemination and use of electronic information.

Although many kinds of communication channels are used for data communication, the importance of the voiceband channel lies in the ubiquity of the switched telephone network and metallic twisted pair subscriber lines. When the need for communication from widely dispersed terminals to host computers became apparent in the 1960s, no widely available data networks with this kind of reach were available. For communicating data between nodes of an airline reservations network, calling in a credit card authorization from a point-of-sale terminal, or transferring data from a bank branch to a central computer, the most economical facilities were analog private lines and particularly switched access lines via the public switched telephone network (PSTN).

Communicating equipment such as terminals and computers produce digital signals designed to travel only a short distance to other equipment nearby. In order to connect equipment separated by longer distances, modems were developed that converted these native digital signals to waveforms capable of passing through analog passband telecommunication channels (Fig. 1). These channels had bandwidth, distortion, and noise characteristics much more accommodating to voice than to data. Telecommunication channels are most often switched, maintaining a fixed connection only for the duration of a call. This is not efficient for the "bursty" activity of a typical data terminal. Despite the later proliferation of specialized data networks with packet-switched communication for efficient sharing of communication links by bursty users, the need for access from locations not directly served by such networks called for voiceband data communication through the public network. The need for higher and higher data communication speeds concurrent with the development of the internet led to the development of dial-up modems that approach the maximum theoretical speed of the band-limited voiceband channel. Millions of modems are sold annually for use with personal computers and computer workstations. These must communicate through the subscriber lines for at least part of the connection to distant enterprise networks via remote access concentrators or internet service providers. Even if the entry node to a specialized data network or virtual private network (VPN) is relatively close to the network end of the subscriber line (Fig. 2), a voiceband data modem or digital subscriber line is needed.

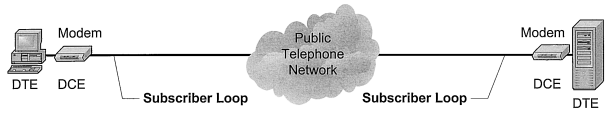


FIGURE 1 Data transmission through a regular telecommunication channel. DTE—Data Terminal Equipment, DCE—Data Communication Equipment.



FIGURE 2 Data transmission over metallic twisted pair subscriber loops for data network access.

Users' equipment varies greatly in data format and speed. As technology and requirements evolve, so do modems. Many different types have been produced, as described in Section III.A; and fundamental distinction is made between asynchronous and synchronous communication. An asynchronous data stream does not have uniformly spaced pulse transitions (where a pulse represents the logic value of a bit or a group of bits); although, as Fig. 3 suggests, it often consists of bursts of uniformly spaced pulses that begin at arbitrary times. A burst may correspond to a particular character key that the user has just pressed. The receiving equipment is synchronized to each character burst, aided by "start" and "stop" indicators. A parity check bit (modulo 2 sum of the bits in a

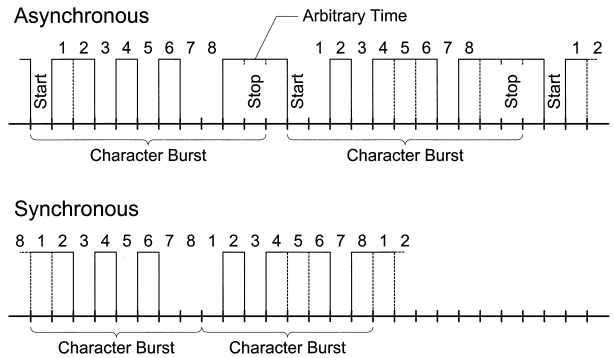


FIGURE 3 Asynchronous and synchronous data streams, illustrated for character burst traffic. In a synchronous data stream, there is no need for start and stop indicators to define timing because of uniform clocking of all pulse transitions.

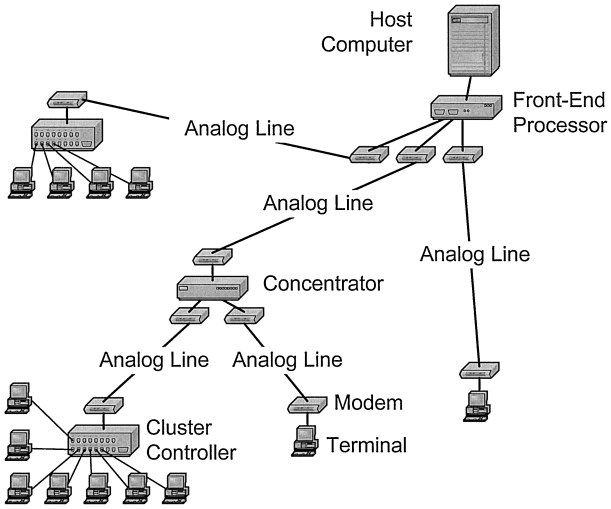


FIGURE 4 Historic inquiry-type data communication network using modems, cluster controllers, concentrators, and a front-end processor.

character representation) can be included in the character burst to allow the receiver to detect (but not correct) some transmission errors. Lower-cost (and lower-speed) terminals are often asynchronous, whereas more expensive (and higher-speed) equipment is synchronous. Asynchronous data communication between terminal and communicating equipment on an individual serial port normally is limited to around 56 or 112 kb/s. The standardization of data network interfaces to synchronous data formats and the availability of lower cost digital transmission and multiplex facilities have led to a broad use of synchronous terminals.

Data communication in commercial applications is often characterized by a network of connections not just a single point-to-point circuit. Figure 4 illustrates the historic network architecture for a private inquiry-type, data communication network that was used in transactional applications, such as a reservations system. The links in this example are all permanently connected, voice grade private lines, leased from local exchange carriers (LECs) and interexchange carriers (IXCs). Cluster controllers coordinate the operation of several terminals at one location so they share the communication link appropriately. Concentrators combine lower-rate data streams into higher-rate streams for further savings on communication lines. Modems at each end of each communication link convert the baseband digital data streams at the terminal equipment interfaces to the continuous passband signals required on the voiceband communication circuits. A front-end communication processor at the host computer manages the network, leaving the host computer free for data processing functions. Although the links in this network are permanently connected, the traffic can be packet-switched if appropriate packet switches and formatters are installed by the network operators (for example, X.25 network with packet assembler–disassemblers, or PADs). Network architectures like this still exist, especially where the terminals are located in remote, rural areas served by analog satellite earth stations or analog microwave transmission systems.

The same network could well include dial-up terminals operating through ordinary switched loops from locations with insufficient traffic to justify a dedicated private line. This type of access is especially common for credit card authorization terminals in stores and restaurants and for access from personal computers and lap-top PCs to office computers and internet service providers for electronic mail (email) and other purposes. In modern applications and networks, the cluster controllers, communication processors, and analog backbone links shown in Fig. 4 have been replaced by remote access concentrators, routers, and digital links as shown in Fig. 5. In the latter, the remote access concentrator accepts incoming dial-up traffic or traffic from dedicated lines and concentrates it for presentation to a LAN router. The router performs authentication and other protocol functions and sends the data out an optical or electrical serial interface port to the digital communication network.

Data traffic, such as characters sent from a keyboard or a bulk file transfer between computers, is digital in origin, but not all data traffic originates in this way. Digital data streams can be derived from analog signals (for example, speech, fax, or voiceband modem signals), transmitted to distant locations with no degradation in quality, and converted back to analog signals. A device called a "CODEC" (COder-DECoder) converts the analog signals

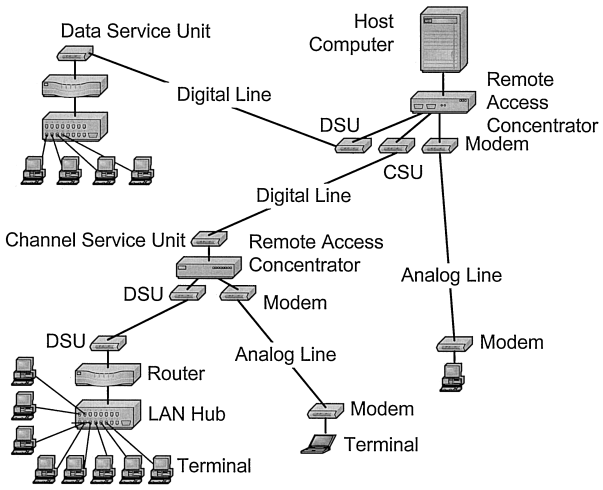


FIGURE 5 Modern digital network associated with inquiry-type, data communication network.

to and from a digital data stream (analog-to-digital, or A/D, and digital-to-analog, or D/A, respectively). Different coding techniques result in different data rates and conversion distortions. In circuit-switched networks (such as the PSTN), 64 kb/s pulse code modulation (PCM) and 16 and 32 kb/s adaptive differential PCM (ADPCM) commonly are used for high-quality telephone speech transmission. Other encoding schemes can compress digitized speech into data streams at rates from less than 2 kb/s to 16 kb/s but these usually have lower recovered quality and include significant processing delays. Nevertheless, these compression techniques are appropriate in many bandwidth-constrained applications, such as digital wireless networks and voice-over-packet applications. Even though the A/D conversion process results in distortion of the original analog signals, there are many advantages of digital transmission. These include (1) nearly perfect regeneration of digital signals regardless of the length of the circuit, (2) flexibility in mixing traffic of different media and service types, (3) ease of encrypting sensitive data, and (4) much more reliable and easily maintained transmission and multiplexing equipment.

Digital circuits carrying digitized analog traffic usually are internal to communication networks and not apparent to the end user. Digital loop carrier systems, access nodes, and primary multiplexers (channel banks) consolidate the outputs of a number of CODECs and combine the digitized analog signals in the telecommunication network. A variety of metallic twisted pair, microwave radio and optical fiber transmission systems, operating at data rates of 1.544 Mb/s (in the United States) and up, carry interleaved streams of digital data and digitized voice.

Although the high-speed links used in these carrier systems require much greater bandwidth than a voice grade channel, they use many of the same signal modulation and channel equalization techniques originally developed for the narrowband telephone circuit. In a typical situation:

- A signal may originate in digital form at a terminal or PC
- Be converted to analog form for transmission through a narrowband analog circuit by a voiceband modem
- Be converted by a CODEC to a digital bit stream representing the analog modem signal
- Be combined with other similar digital signals in a multiplexer to form a higher-speed aggregate signal within the network for transmission on a digital carrier system
- Be demultiplexed and converted back to analog form for delivery to the other end of the channel by another CODEC
- Be converted by another modem to a replica of the original digital data signal

The telecommunication networks of industrialized countries are evolving toward all-digital, multimedia networks that will eventually provide end-to-end carriage of data, digitized voice, and digitized video traffic. Some services involve integrated handling of different media and extensive user signaling to and through the network to setup a variety of calling and receiving options. The integrated services digital network (ISDN), a concept for a family of standard digital interfaces and services and having compatibility with existing analog and digital networks, continues to evolve. Voiceband data communication may eventually be superceded by end-to-end digital communication via an ISDN, but the principles of data communication developed for analog voice grade circuits will continue to find applications on the links and subscriber lines embedded in digital networks.

The transmission of information between a computer and other computers or the internet through the voiceband channel is a familiar application of modems. In addition to these familiar applications, low-speed voiceband modems, using standard protocols, are used to send caller number identification (CID or CND) and caller name (CNAM) from the central office switching system to a subscriber who subscribes to these features. The calling information is sent through the public telephone network (as in CID) or determined from a database embedded in the network (as in CNAM) and then transmitted to a box or telephone with a caller ID display on the called party's premises. These subscriber features are two of many Custom Local Area Signaling Services (CLASS®) offered via all modern switching systems.[3]

II. VOICEBAND CHANNELS

The focus of this section is on the analog voiceband channel, the one to which the telephone subscriber is directly connected and the channel that is often used for end-links in private data networks. This channel also is the one for which most of the advances in data communications and signal processing have been made. The voiceband channel is not a single or unique entity, but rather a family of types and characteristics, as this section will describe as a prelude to the introduction of modems in Section III.

The analog voiceband channel seldom is a physically separate wireline or dedicated wireless circuit from end to end. If the channel is a long-distance circuit, it usually is a tandem connection of a number of different transmission facilities, mostly digital but possibly some analog. As the last section pointed out, some of these may be carrier facilities. Modern telecommunication networks

[3]CLASS® is a registered mark of Telcordia.

are based on time division multiplexed (TDM) digital carrier facilities consisting of optical transmission systems, such as Synchronous Optical Network (SONET), with traditional circuit-switching systems and Asynchronous Transfer Mode (ATM) packet-switching and traffic control technologies (Fig. 6). A carrier system is a bulk transmission system between network operators' central offices or between an access node and a central office. A voice channel entering a carrier system passes through a line unit or channel unit (subscriber line interface circuit, or SLIC), which processes the signal in some appropriate way (for example, A/D and D/A conversion) and multiplexes it onto high-speed transport circuits in the carrier system. The analog-to-digital coding initially takes place using 64 kb/s PCM (8000 samples per second with each signal sample converted to an 8-bit data word). Further processing using ADPCM or other speech processing techniques may reduce the bit-rate to some lower value as previously mentioned. Such processing introduces additional nonlinear distortions and other distortions into the recovered signals.

A carrier system may be connected in tandem with other network carrier systems and finally with a destination subscriber circuit, usually a metallic twisted pair. It is useful to know how a voiceband circuit is put together when investigating the sources of channel impairments, but the user

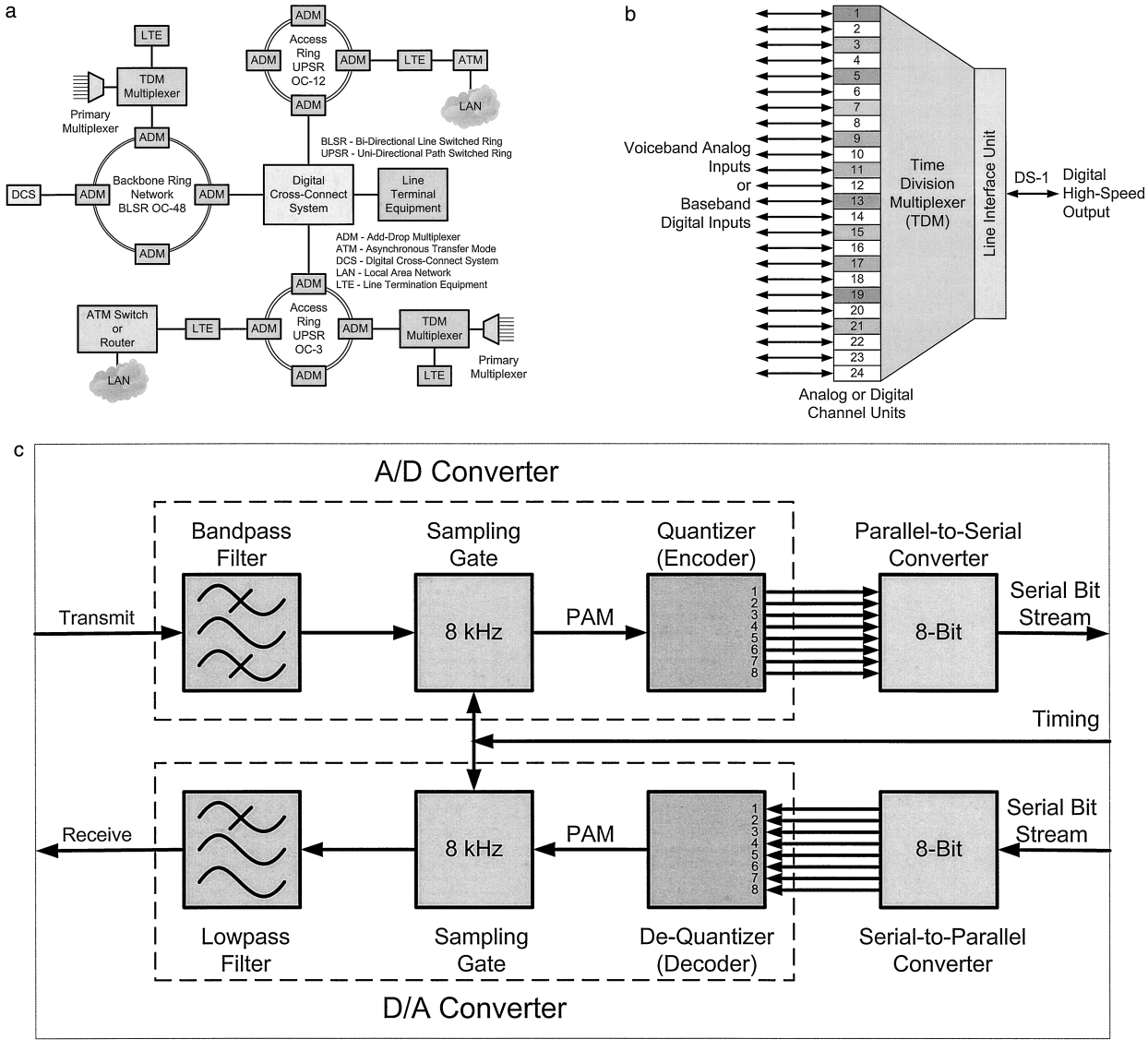


FIGURE 6 Carrier transmission system. (a) Fiber-optic transmission system based on Synchronous Optical Network (SONET). (b) Primary multiplexer for 24 analog voiceband channels. (c) Analog channel unit block diagram.

usually will be satisfied to think of the circuit as a black box with specified characteristics. These characteristics will be described later in this section.

The analog voiceband channel is available in a number of forms. Reference already has been made to the private line, which usually is a four-wire circuit in which separate two-wire (or equivalent) transmission channels are provided for each of the two directions of transmission (Fig. 7a). By definition, a private line is not switched in the public network. It is referred to as a *special access* circuit in modern vernacular. On the other hand, circuits that are switched are referred to as *switched access* circuits. Data communications usually are full-duplex (that is, simultaneous bidirectional communication). Transmission within the modern telecommunications network is, in fact, fundamentally full-duplex, with separate channels in each direction. Therefore, all that is required to complete a four-wire circuit to a subscriber is two, two-wire subscriber lines. Alternately, a device called a hybrid transformer (or electronic equivalent) can be used to convert the four-wire circuit to a bidirectional two-way circuit for final connection to the subscriber. "Virtual" private lines, which appear to the user as hard-wired circuits but may actually consist of logical connections using different facilities or paths at different times, are commonplace in modern networks. Frame relay and ATM are two such technologies that route multiple data traffic logical connections over a single physical facility or circuit.

An extension to multiple locations is offered in the multipoint circuit (Fig. 7b). This is a four-wire special access circuit used in polling applications, where a central processor interrogates in turn a number of terminals. By connecting many terminals to one line, the transmission costs are considerably lower than what they would be if each terminal had an individual connection to the central processor. The connection points are bridge circuits that match impedances and provide signal level control. The bridge circuits may or may not have control and diagnostic capabilities such as the capability to switch a particular terminal in or out. Much attention has been given in data communication research to efficient operation in polling applications, especially quick adaptation of the channel equalizer in the receiver of the central polling modem as it jumps from listening to one distant transmitter to another.

Special access circuits are attractive to businesses, internet service providers, and others that have enough traffic to justify the monthly costs. They have the important advantage of full-time connectivity, without the set-up time of a switched access connection.

A dial-up line is a two-wire circuit (communication in both directions on the same pair), but the circuit is switched on a four-wire basis and transmitted on four-wire equivalent transmission facilities. The advantages of a dial-up lines are their availability from any subscriber location and long-distance charges, if any, that are paid only for the actual time of use. Modem dial-up lines are appropriate for individuals at home and small businesses or any other location with relatively low traffic or where more advanced technologies, such as xDSL, are unavailable or unaffordable.

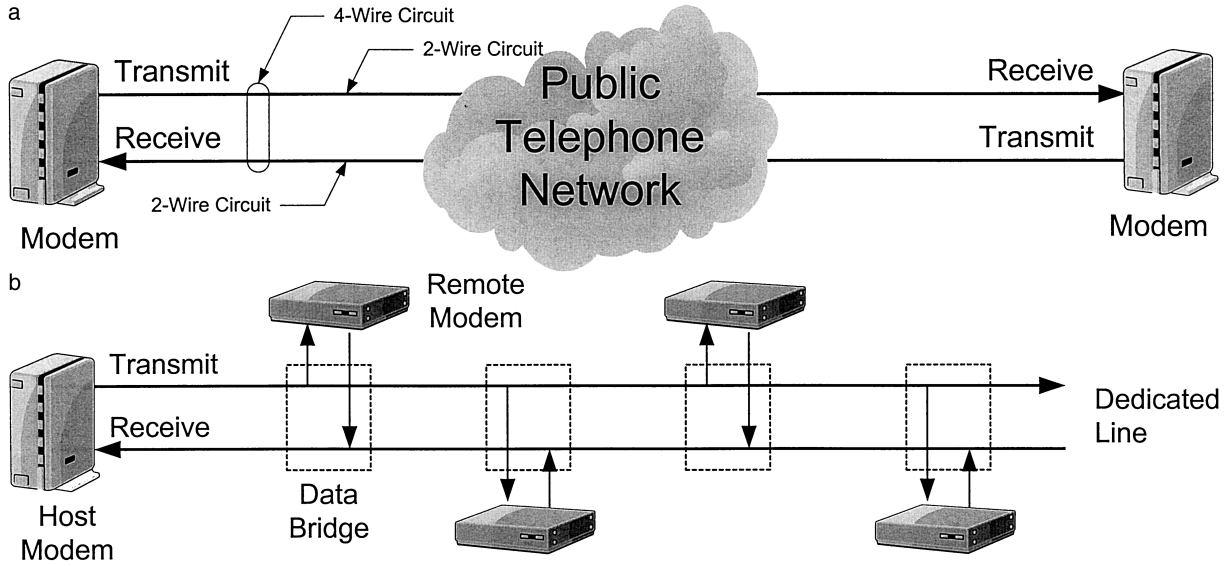


FIGURE 7 Private lines. (a) A four-wire private line, or special access circuit, is a dedicated connection through the network with a separate channel for each transmission direction. (b) A multipoint circuit supports geographically separated terminals polled by a host computer. This type of circuit is used only where lopsided tariffs encourage the use of analog modems.

Unless the terminals connected to the dial-up line are operated simplex (one direction only) or half-duplex (one direction at a time), the signals traveling in the two directions must be separated by the modems to prevent self-interference of a receiver by its own transmitter. Full-duplex data transmission (Fig. 8) is achieved either by frequency separation, in which the two transmission directions are in separate frequency bands (a form of frequency division multiplexing, or FDM), or by echo cancellation, described in more detail later, in which the same frequency band is used for both directions and self-interference is adaptively cancelled. Some modems use both methods.

The conversion from a two-wire to four-wire circuit is carried out in the central office line circuit, channel unit or two-wire modem by a device called a hybrid transformer (also called hybrid coupler or, simply, hybrid), shown in Fig. 8. This is a passive or active four-port device that functions as a directional coupler, theoretically preventing transmit direction signals from reflecting back into the receive direction. In practice, impedance imbalances lead to some leakage (noninfinite trans-hybrid loss) and hence to returning echo. These are not harmful in a transmission scheme using separate frequency bands for the two transmission directions (Fig. 8a) but must be effectively cancelled when both directions use the same bandwidth (Fig. 8b).

The actual routing of a call from one local exchange area to another is shown in Fig. 9. The call is handed off from the local exchange carrier (LEC) network to a transit network (interexchange carrier, or IXC) connecting each exchange area. The LEC and IXC networks may be owned and operated by the same network operator and are differentiated here only for convenience. The handoff from one network to another is transparent to the user, but the delay and distortion characteristics of the connection may change if any intermediate analog transmission facilities are involved. When all-digital transmission and switching facilities are used, there is no additional distortion unless digital circuit multiplication equipment (DCME) is used. Such equipment normally is used on very long (and expensive) transmission facilities such as undersea and satellite transmission systems to reduce speech circuit bandwidth requirements.

The main transmission characteristics of an analog channel can be represented by its attenuation and phase delay. The twisted pair portion of a typical channel is linear

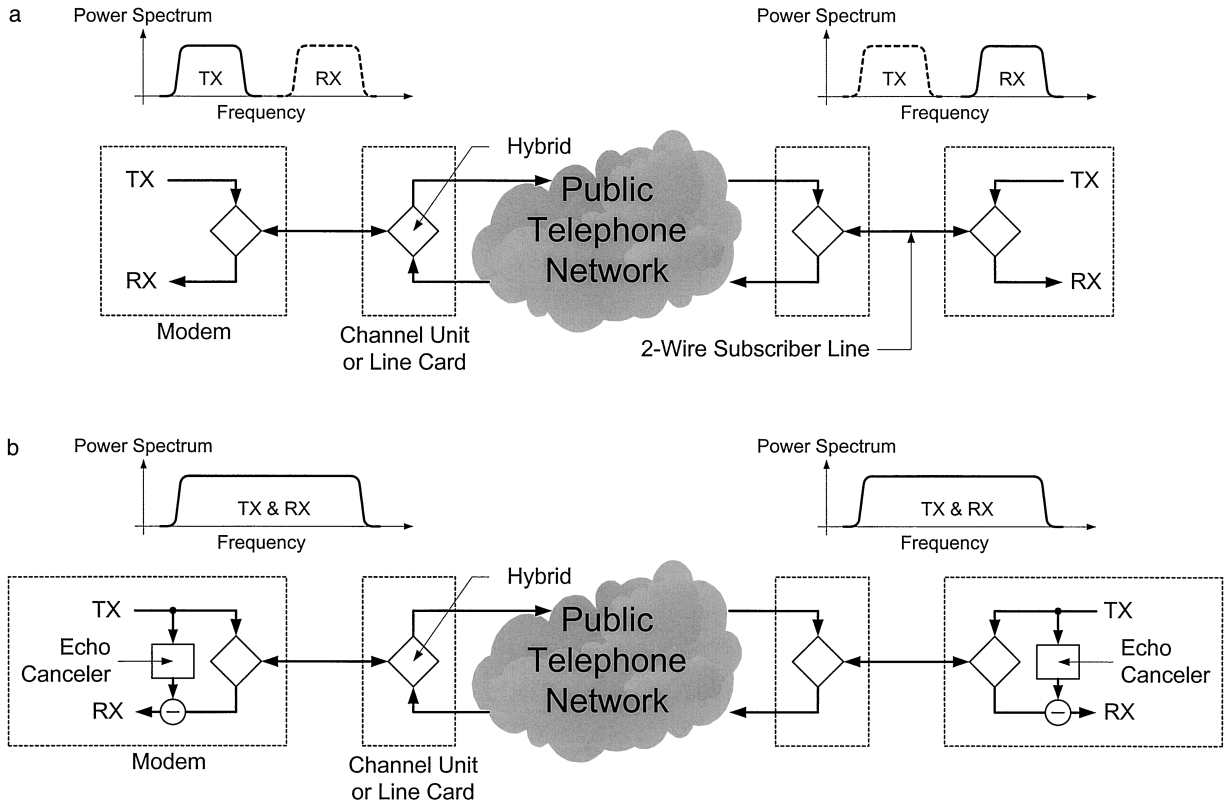


FIGURE 8 Full-duplex data communication over dial-up telephone lines. (a) Separate frequency band (a little less than one-half the channel) in each transmission direction. (b) Use of the same full-channel frequency band in both directions, with echo cancellation to suppress self-interference.

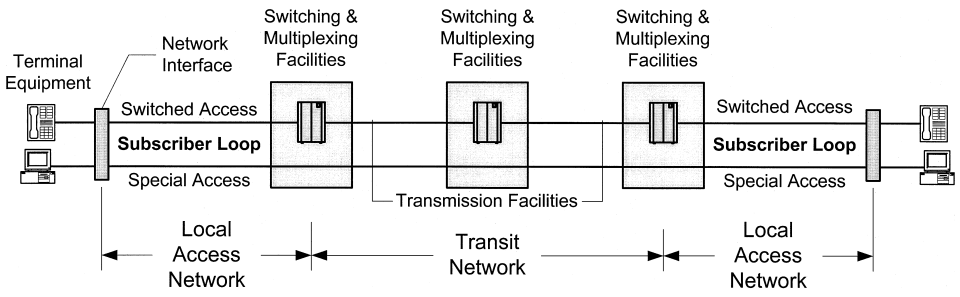


FIGURE 9 Local and transit networks.

(a particular input results in a single output) and only introduces linear distortion (such as amplitude and phase distortion) due to its inherent transmission characteristics. Such distortions take place at lower and higher frequencies where these characteristics change. The line circuits or channel units to which the twisted pairs are connected introduce nonlinear distortion (a particular input results in more than one output), but the resulting nonlinear distortion is controlled. Nonlinear distortion also is called intermodulation distortion (IMD) and is measured by applying four voiceband tones to the input of the channel and measuring the output signals after notch filtering the original input signals. No output signal will be measured in a linear channel. A nonlinear channel will have a measurable output signal.[4]

Figure 10(a) shows amplitude measurements taken between two analog channel units of a digital loop carrier (DLC) system. The measurements do not include the effects of any subscriber loop cable. Note that the channel has a passband characteristic, with decreasing transmission below 200 Hz and above 3400 Hz. Although the phase delay is a component of the channel's transmission characteristic, the change in delay with respect to frequency, called envelope delay distortion, is more important to modem transmission than phase delay itself. The envelope delay distortion is related to the phase delay $\phi(\omega)$ by the derivative

$$EDD(\omega) = -d\phi(\omega)/d\omega,$$

where $\omega = 2\pi f$ is the angular frequency, measured in radians per second, and f is frequency in Hertz. The envelope delay distortion expresses the dispersion of the channel through relative differences between the transmission delays of different spectral components. Both the amplitude and delay distortions contribute to the smearing of signal pulses and the consequent intersymbol interference (ISI) that must be corrected in higher-speed modems. Figure 10(b) shows the envelope delay distortion of the same channel represented in Fig. 10(a). An actual end-to-end channel includes the effects of subscriber loops connected to the channel units. Subscriber loops using metallic twisted pair cables have a lowpass characteristic in which attenuation increases with frequency. Figure 10(c) shows an end-to-end channel consisting of one fairly long subscriber loop (9.7 dB loss at 1004 Hz) at one end and a zero length loop connected to an ISDN terminal adapter at the other end (the TA introduces 8.1 dB of loss at 1004 Hz). Note that the frequency rolloff at low frequencies is due entirely to the filters in the A/D and D/A converters. High-frequency rolloff is a combination of the filters in the A/D and D/A converters and the characteristics of the cable. Although the EDD of the end-to-end connection is not shown here, it would be dominated by the EDD of the DLC channel units.

It is usual to neglect the absolute delay, which would be the constant part of $EDD(\omega)$, and consider just the relative delay of different spectral components. Although absolute delay is neglected in this context, it is nonetheless of great importance, especially in the transmission of speech and video and the affects such delay has on human interaction and on the operation of echo cancelers. Signal processing delays, and delays caused by digital network elements, usually are much larger than the delays introduced by end-links such as the subscriber loop. Subscriber loop lengths, and thus their propagation delays, are short. Long-distance connections, such as intercontinental connections and especially connections over geostationary satellite circuits, introduce significant delay, which adds to the delay of any digital network elements (switches, digital cross-connect systems, speech processing devices, and multiplexers).

Linear distortion arises from the practical limitations of network elements, particularly CODEC filters (Fig. 6). These filters are designed to limit the spectra of voice and other analog signals to avoid low-frequency noise and aliasing in analog-to-digital converters. Such filters were not originally intended to minimize phase distortion in the passband, since phase distortion normally does not

[4]The output signal from a nonlinear channel typically will have numerous frequency components but the second- and third-order products are the most significant. For input signal frequencies A and B, the important second-order products are B − A and B + A and the important third-order product is 2B − A.

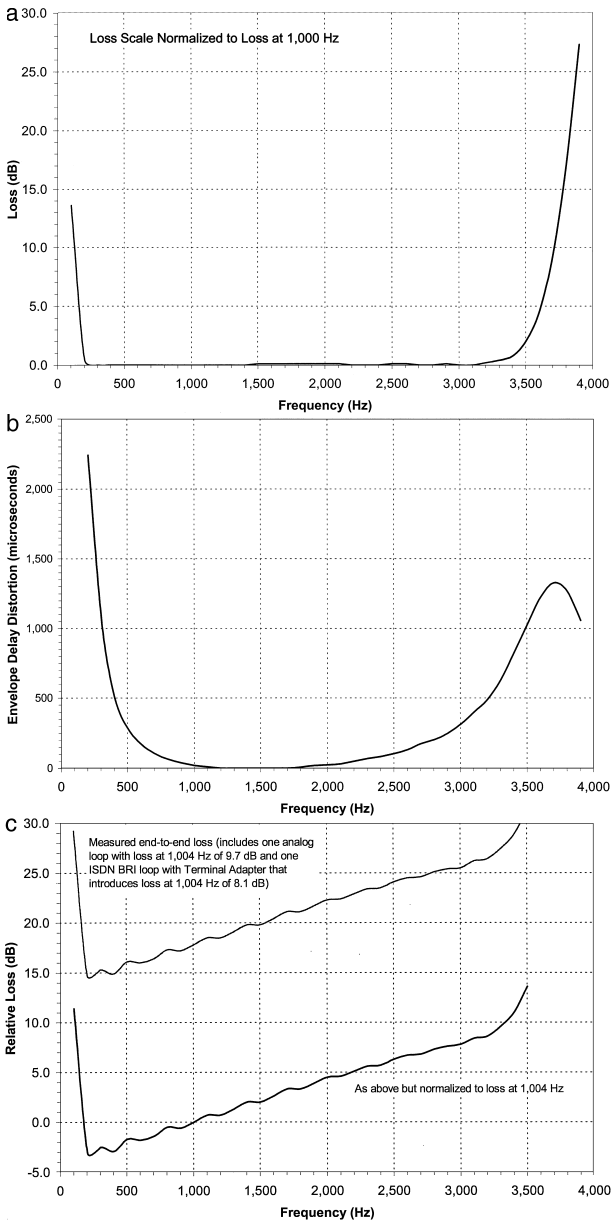


FIGURE 10 Typical attenuation and envelope delay distortion of a voiceband telecommunication channel. (a) DLC channel loss, (b) DLC EDD, (c) end-to-end channel loss.

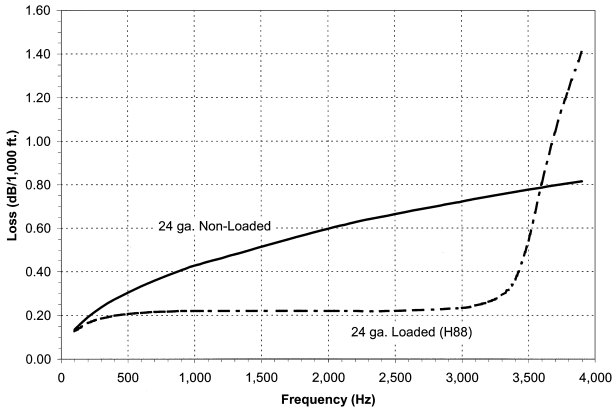


FIGURE 11 Comparison of loaded and nonloaded twisted pair transmission characteristics.

degrade analog speech signals. The distortions introduced by twisted pair cables depend on their length and composition and whether they are loaded or nonloaded. A loaded cable consists of twisted pairs with lumped inductors inserted in series with the pairs at fixed intervals.[5] Loading was used to compensate for the capacitive reactance of the twisted pairs and, thus, to improve the voiceband frequency attenuation on loops longer than 18,000 ft (loading on these loops reduces the transmission loss per unit distance at 1000 Hz by about one-half and at 3000 Hz by about one-third). However, loading greatly increases the attenuation (loss) and envelope delay distortion at higher frequencies. This was of no consequence to speech transmission but is a serious impediment to voiceband data transmission. Although there is a huge embedded base of loaded cables in today's network, their use in modern plant construction is small because loaded plant is not compatible with digital subscriber line technologies. Figure 11 compares the loss-per-unit distance of a nonloaded and loaded 24 ga. twisted pair.

The second most prominent characteristic of a voiceband telecommunication channel is noise. It is the sum of many small contributions of thermal noises, powerline induction, radio noises, atmospheric electrical discharges, crosstalk interference, and other products of natural and man-made processes (Fig. 12). Although there can be occasional large noise spikes (impulse noise), the sum noise usually is well modeled as Gaussian noise, which is analytically convenient. The noise usually is at a low level compared to signal strength. An important criterion is the ratio of the data signal power at the input of a modem receiver to the noise power present at that point (signal-to-noise ratio, or SNR). The effects of noise on modem receiver operation is entirely different from its effect on human speech communication. In the latter, the actual noise amplitude, and not SNR, is important because relatively large noise values are unnoticed during active speech but relatively small amounts of noise may be annoying during speech pauses. Twisted pairs are specified to have a measured noise power no greater than −70 dBm (+20 dBrn) when measured with a filter having a C-message characteristic.[6] This filter takes into account the combined effects of

[5]Two loading schemes are most common, D66 using 66 mH inductors at 4500-ft spacing and H88 using 88-mH inductors at 6000-ft spacing.

[6]dBrn is deciBels with respect to reference noise; reference noise is 1 pW, or −90 dBm.

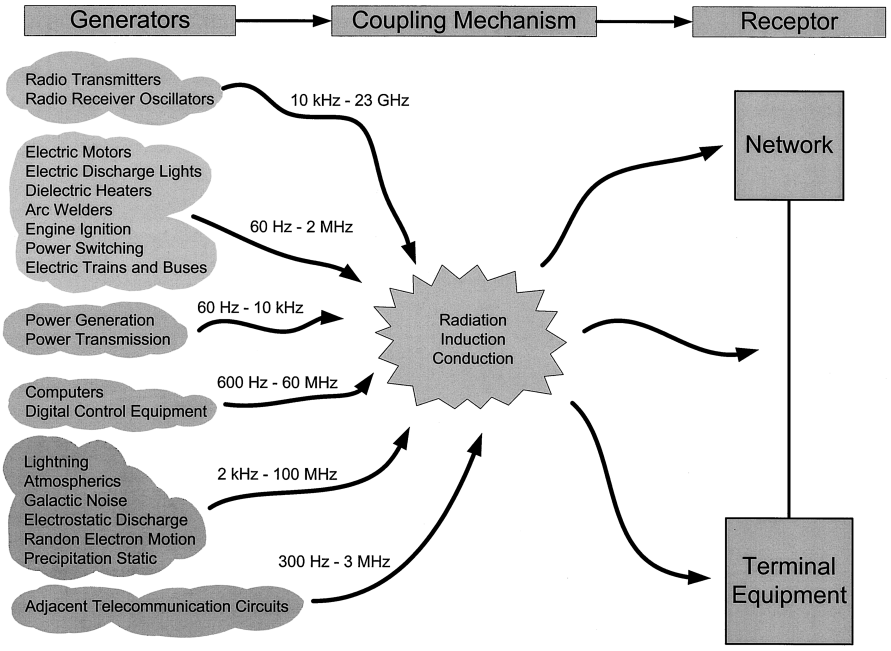


FIGURE 12 Noise coupling (from intro course).

human hearing and the telephone set frequency response. In the voiceband, the predominant noise is due to powerline induction at fundamental and harmonic powerline frequencies.

So-called "toll quality" telecommunication channels are considered to have an SNR for speech purposes of at least 30 dB. The noise bandwidth of a voiceband modem receiver input filter is around 3.4 kHz. Assuming the rest of the channel does not narrow this bandwidth any further, this allows for a maximum possible error-free information transfer rate (channel capacity) of almost 34 kb/s. This value is determined from the well-known formula of information theory (known as the Shannon-Hartley theorem)[7]

$$C = W \log_2(1 + S/N), \tag{1}$$

where S is the signal level in watts, N is the average white (uniform-spectrum) Gaussian noise power in watts (WN_0, where N_0 is the spectral power density in watts per Hertz) in bandwidth W Hz ($S/N =$ SNR), and C is the maximum channel capacity in bits per second. Equation (1) predicts a maximum spectral efficiency of 10 b/s/Hz of bandwidth for SNR $= 30$ dB. The derivation of this formula provides additional insight into the signaling problem, in particular, the result that power should not be applied uniformly across the frequency spectrum passed by the channel, but instead more heavily at those frequencies that the channel attenuates the least.

[7]Equivalently, $C = 2W \log_2 \sqrt{1 + S/N}$. For computational simplicity, note that $\log_2 x = \log_{10} x / \log_{10} 2$.

In practice, it takes a complex coding and decoding effort to approach the calculated 34 kb/s line speed. It is important in the design of modems to avoid or minimize noise enhancement in the process of compensating for other impairments. In the end, it can turn out that the other significant channel impairments described in the following can be the most serious constraints on performance at higher speeds.

One of the additional impairments is perturbation of carrier phase. Phase jitter and frequency offset (or phase roll) are associated with the oscillators used for frequency translation of voice channels into frequency slots in frequency division multiplexed (FDM) carrier systems. Fortunately, FDM carrier systems are little used in the modern public network. This source of voiceband channel impairment is negligible in digital carrier systems.

Another impairment is nonlinear distortion (also called intermodulation distortion), for which compensation is difficult. It can sometimes be modeled (Fig. 13) as a memoryless (present signals are not affected by past signals) third-order polynomial function embedded between linear distortions. The level of second- and third-order

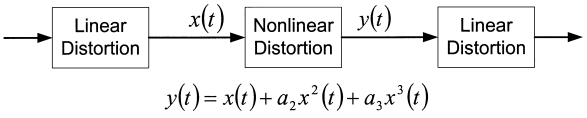


FIGURE 13 Simple model for channel with both linear and nonlinear distortions (noise and phase perturbations not shown).

TABLE I Modem Susceptibility to Transient Impairments [ANSI T1.512-1994]

Modem type (1)	Impulse noise level (dBrnC) None cause error (2)	Impulse noise level (dBrnC) All cause error (3)	Minimum phase hit (Degrees at 1004 Hz) (4)	Modem speed (b/s)
V.22	74	80	18.7	1,200
V.22bis	67	72	7.0	2,400
V.26	63	66	25.0	2,400
V.27	59	63	12.5	4,800
V.32	59	63	8.0	9,600
V.32bis	59	57	4.0	14,400

spectral components are typically 30 to 40 dB below the fundamental. Nonlinear distortion is introduced by the analog-to-digital conversion process and digital attenuators (loss pads) in line cards and channel units. These attenuators introduce a controlled amount of loss in the digitized received signal for transmission purposes but have the effect of introducing nonlinear distortion manifested as a higher noise level. Another source of nonlinear distortion is the low bit-rate encoding methods (for example, ADPCM) used in some transmission equipment. These methods originally were designed for speech transmission and can severely impact voiceband data transmission. Nonlinear distortion can be the limiting factor in high-speed voiceband data communication. Considerable data have been accumulated characterizing voiceband telecommunication channels and the performance of most types modems through them (Table I). Unfortunately, no similar data have been publicly published for the higher-speed V.34, V.90, and V.92 modems. Table I may be interpreted as follows: The impulse noise levels shown are thresholds. If the impulse noise is below the threshold given in Column 2, there will be no errors caused by the impulses. If the impulse noise is above the threshold given in Column 3, there will be errors caused by all impulses. The phase hits (abrupt phase changes) above those given in Column 4 will cause errors.

There are additional abrupt disturbances—for example, impulse noise, phase and gain hits, and dropouts—on telecommunication circuits stemming from the peculiarities and operating environments of various transmission and switching systems. These usually are constrained by telecommunication equipment and system design.

If trunking and transmission facilities are all-digital, the performance of switched access channels are determined by the subscriber lines connected to the modems. There is quite a variation in loop performance because of the variation in loop length, bridged tap length, and exposure to powerline induction and impulse noise. The basic requirements for switched access loops are given in Table II. Circuit noise is measured across the balanced twisted pair while powerline induction is measured longitudinally to ground. The circuit balance parameter indicates how well the twisted pair is balanced with respect to ground and, thus, its ability to convert longitudinal noise (powerline induction) to circuit noise.

Analog voice grade special access channels are designed to have specific transmission characteristics with little variation. These are more detailed than switched access loops and are specified in industry technical references, such as the VG-series published by Telcordia [TR-TSY-000333] and in standards published by American National Standards Institute [ANSI T1.506-1997] and [ANSI T1.512-1994]. As an example, the amplitude and envelope delay distortion limits for the VG6 circuit are shown in Fig. 14. Table III summarizes other VG6 transmission requirements.

While the Telcordia voice grade requirements are widely used, they do not represent industry consensus and they reflect the limitations of the legacy network prevalent in the Bell Operating Companies (BOC), for which these requirements were written. The ANSI standards are more widely accepted as true industry standards (but they also recognize network limitations). In order to accommodate a wide range of users, the ANSI standard referred to here [ANSI T1.512-1994] specifies three groups of performance objectives—Group 1, 2, and 3. Group 1 is intended for use with high-speed voiceband modems and is the most stringent, followed by Group 2. Group 3 is the least stringent and applies to analog networks or digital networks with significant analog network elements. Table IV

TABLE II Switched Access Loop Transmission Performance

Transmission parameter	Limit
Loop loss	≤8.0 dB
Circuit noise	≤20 dBrnC
Powerline induction	≤80 dBrnC
Circuit balance	≥60 dB

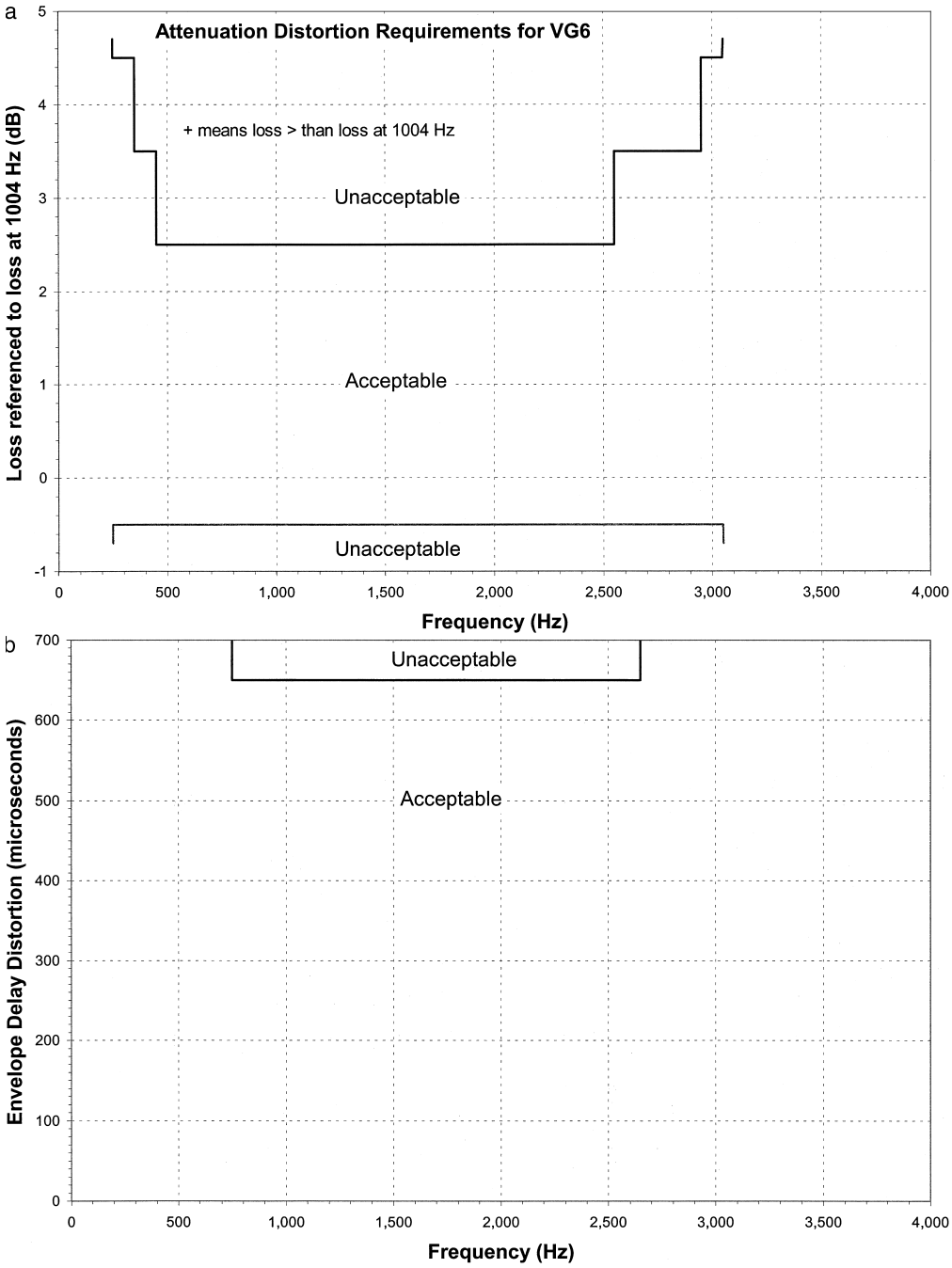


FIGURE 14 Specified amplitude distortion and envelope delay distortion characteristics for a VG6 analog voice grade special access channel.

summarizes these objectives. Phase jitter and frequency offset are not specified for Groups 1 and 2 because digital facilities do not generate phase modulation impairments such as these to any significant degree. The predominant contribution of phase jitter is from quantization noise in analog-to-digital converters. Transient impairments, including phase and gain hits and dropouts, are not specified because they are difficult to measure accurately and some measuring sets give inconsistent results. These transients usually are caused by timing slips and protection switching in digital networks. Performance specifications include acceptance limits for new circuits and immediate action limits for in-service circuits. Test methods and equipment for analog voiceband circuits are standardized,

TABLE III VG6 Transmission Requirements [TR-TSY-000333]

Transmission parameter	Acceptance limit	Test conditions
Attenuation distortion	See Fig. 17	Swept tone at −13 dBm0
Envelope delay distortion	See Fig. 17	As specified in [IEEE Std. 743-1995]
C-message noise	≤27 dBrnC0	Channel terminated in its rated impedance
C-notched noise	≤45 dBrnC0	−13 dBm0, 1004-Hz holding tone
Signal-to-C-notched noise ratio (SCNN)	≥32 dB	−13 dBm0, 1004-Hz holding tone
Notched impulse noise	≤15 counts in 15 min	Threshold set to 65 dBrnC0, 1004-Hz holding tone
Phase jitter (20- to 300-Hz filter bandwidth)	≤4.0° Peak–peak	−13 dBm0, 1004-Hz holding tone
Phase jitter (4- to 300-Hz filter bandwidth)	≤9.0° Peak–peak	−13 dBm0, 1004-Hz holding tone
Second-order intermodulation distortion (R2)	≥35 dB	4-tone
Third-order intermodulation distortion (R3)	≥42 dB	4-tone

leading to consistent results for most (nontransient) measurements [IEEE Std. 743-1995].

The combined effects of noise, jitter, and distortion on a channel can be viewed on an oscilloscope by observing the eye pattern (Fig. 15). The eye pattern is the superposition of signals representing all possible pulse sequences just before the detector (decision device) in the data receiver. For a binary system, the eye pattern indeed looks like an eye, which closes if noise and distortion are too large. The more open the eye, the greater the margin of safety against errors. The best sampling time is where the eye is widely open and a small error in the sampling time does not lead to a significantly smaller safety margin. Automatic adjustment circuits are used to adjust the sampling time, sometimes in conjunction with a data scrambling system to guarantee an adequate number of data level transitions for clock recovery.

Eye closure at the sampling time can be quantified by an analytical expression such as mean-squared error (MSE), which is the expected value of the squared difference between a sample value and the signal level expected on an ideal (noiseless and distortionless) channel. Most channel equalizers (described in Section V) that reduce intersymbol interference and thereby improve the eye pattern are adapted by algorithms seeking to minimize the MSE.

TABLE IV Transmission Objectives for Point-to-Point Voiceband Data Special Access Circuits [ANSI T1.512-1994]

Transmission parameter	Group 1	Group 2		Group 3
		Tier 1	Tier 2	
Loss deviation (dB)	±1	±1	±1	±1
Attenuation distortion (dB)				
404–2804 Hz	−2.0/+5.0	−0.8/+1.5	−0.8/+3.5	−1.5/+9.5
304–3004 Hz	−2.0/+6.0	−0.8/+1.5	−0.8/+4.5	−2.5/+11.5
304–3204 Hz	−2.0/+8.0	Not specified	Not specified	Not specified
Envelope delay distortion (μs)				
804–2604 Hz	≤550	≤150	≤650	≤1,700
604–2804 Hz	≤800	≤200	Not specified	Not specified
504–2804 Hz	Not specified	≤550	Not specified	Not specified
504–3004 Hz	≤1,500	≤2,950	Not specified	Not specified
Intermodulation distortion (dB)				
R2	≥49	≥32	≥32	≥26
R3	≥50	≥42	≥42	≥34
Signal-to-C-notched noise ratio (dB)	≥33	≥32	≥32	≥26
Notched impulse noise Threshold (dBrnC)	63	65	65	69
Phase jitter (° peak–peak)				
20–300 Hz	Not applicable	Not applicable	Not applicable	≤8
4–300 Hz	Not applicable	Not applicable	Not applicable	≤13
Frequency offset (Hz)	Not applicable	Not applicable	Not applicable	±1

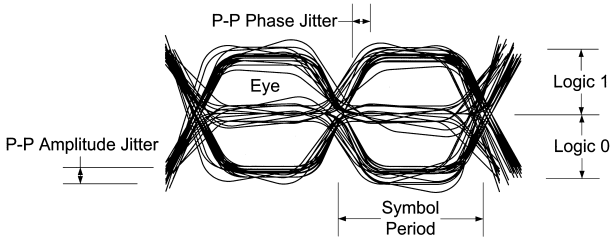


FIGURE 15 Eye pattern of a binary antipodal pulse train over two symbol intervals. As intersymbol interference increases, the eye closes making detection of the correct symbol more difficult.

To communicate full-duplex data, steps must be taken to disable certain network elements that aid voice communication but interfere with data communication. Long-distance telephone channels were equipped with echo suppressors starting in the 1920s to prevent the return of disturbing echoes to a speaking party. On the assumption that only one party at a time is likely to be talking, an echo suppressor blocks a signal from the other end of the line in addition to the undesired echoes. Simultaneous two-way conversation, be it speech or data, is prevented. Starting in the 1980s, echo suppressors were replaced by echo cancelers in modern networks. Echo cancelers selectively eliminate most local and distant echoes without blocking the signal from the other party. Because of modem adaptation requirements, however, both echo suppressors and echo cancelers cause trouble for full-duplex data transmission. Echo cancelers are initially disabled by specified signaling (a voiceband tone with specific frequency and phase characteristics) and kept disabled by transmission of a minimum signal power in a specified frequency band.

III. MODEMS

Modems already have been defined as the translators between digital terminals and analog transmission circuits (Fig. 1). Voiceband modems may be separate units or built into equipment (in personal computer parlance, external or internal modems). They may be half- or full-duplex, used on private or dialup lines, and operate at data rates from 300 to as high as 56 kb/s.

For a given data rate, the modem must simultaneously achieve several conflicting objectives. The *first objective* is high spectral efficiency, measured in b/s per Hz of bandwidth. For example, with a nominal channel bandwidth of 2400 Hz, the spectral efficiency varies from 1 b/s Hz for a 2.4 kb/s modem to 6 b/s Hz for a 14.4 kb/s modem. If higher bandwidth is available, say 3400 Hz, practical values of efficiency are 8 to 10 b/s Hz. High spectral efficiency calls for modulation techniques that trade off either lowered noise immunity associated with larger signal sets or increased channel distortion from placing signal energy in the distorted edges of the transmission band, or both. Bandwidth normalized spectral efficiency versus SNR is plotted in Fig. 16 (this is a plot of Eq. 1 using $W = 1$ Hz).

The *second objective* is low bit error rate (BER). BER is the statistical expectation of the proportion of data bits at the receiver output that are incorrect. In many applications, the BER is considered high if it exceeds 10E-5 (one

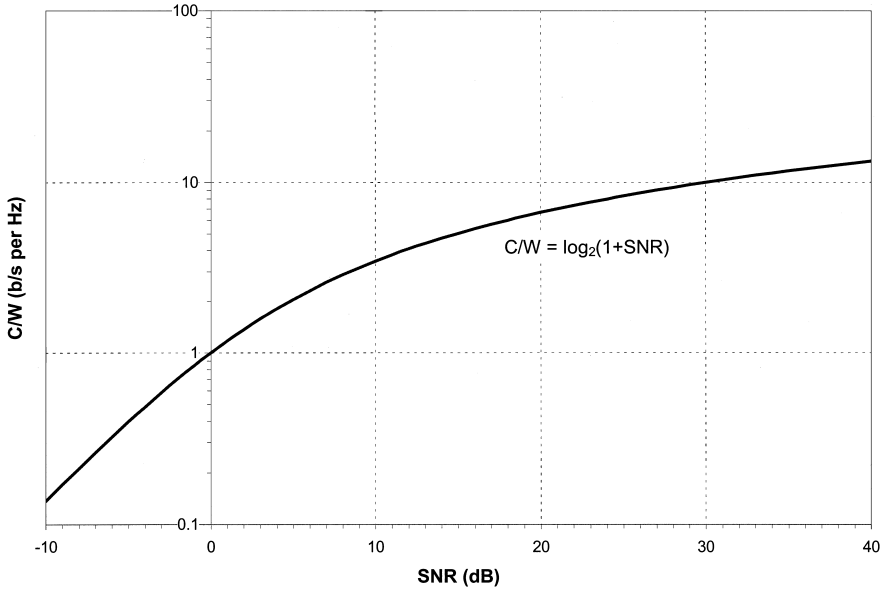


FIGURE 16 Normalized channel capacity vs. channel SNR.

error in 100,000 bits transmitted). Data communication technology has made great strides in increasing spectral efficiency while maintaining an acceptably low BER. A *third objective*, for higher-speed modems that must adaptively equalize, or correct, the transmission channel, is quick adaptation to changes in the channel or changes to different channels. The latter occurs in calls from one location to many different locations through the PSTN. Rapid adaptation, like spectral efficiency, conflicts with good steady-state error rate performance. A *fourth objective*, perhaps the most important, is low cost for a given level of performance. VLSI circuits have made possible dramatic reductions in modem costs and increases in their computational sophistication.

Figure 17 illustrates the architecture of an asynchronous modem, which generates and recovers the clock from the data using start and stop indicators. This illustration does not show ancillary functions such as DTMF and tone generation and accompanying multiplexer. In voiceband modems, the baseband digital data presented at the user interface modulates a carrier by changing the carrier amplitude, phase or both. Modulation results in symbols separated by a predetermined interval. The symbols may overlap, meaning each symbol may extend over more than one symbol interval causing intersymbol interference.

The modulated carrier is a passband (voiceband) line signal. The receiver's demodulator recovers the original baseband signal and presents it to the user interface. Frequency, phase, and amplitude modulation, or a combination, are commonly used (Section VI). Line signal level coding and modulation techniques extending over several symbol intervals can be exploited to achieve better performance for a given SNR. Phase perturbations in the received signal can be tracked by an adaptive local oscillator in the receiver. The correct sampling time also can be selected by the adaptive control circuitry.

The heart of a higher-speed modem is its channel equalizer (Section V), which is an adaptive linear filter that largely corrects for the linear distortion of the transmission channel. In many channels, the linear distortion causes severe intersymbol interference among pulses. At rates of 4800 b/s and higher, channels that yield totally unacceptable error-rate performance without equalization are conventionally equalized to provide BER of 10E-6 and lower. Slow changes in linear channel characteristics are easily tracked by such an equalizer. Advances in VLSI technology have made possible very high-speed digital signal processing (DSP) integrated circuits of relatively low cost and of high computation capacity for equalization, filtering, phase tracking, and other modem functions.

A number of equalizer structures have been developed, but the most common are the tapped delay line (also called Finite Impulse Response, or FIR, or transversal), filter, and the decision-feedback equalizer (DFE). In both cases, decisions made on the equalized signal and resulting data stream are used to adapt the tap weights (filter coefficients), which are the variable elements of the equalizer. With DFE, once the value of the current symbol is determined, the intersymbol interference contribution of that symbol to future received symbols can be removed. An alternative approach, which also can be combined with an equalizer, is to compute the most likely sequence of symbols transmitted in a data communication session, based on observation of the noisy and mutually interfering data pulses received from the channel.

In order to minimize intersymbol interference, the transmitted signals from modems are shaped. Signals providing the shape of a sinc function exhibit no ISI but such a function requires an ideal filter. A raised cosine pulse shaping filter provides a compromise and is used in modern modems. Such a filter is implemented as a finite impulse response (FIR) filter using DSP.

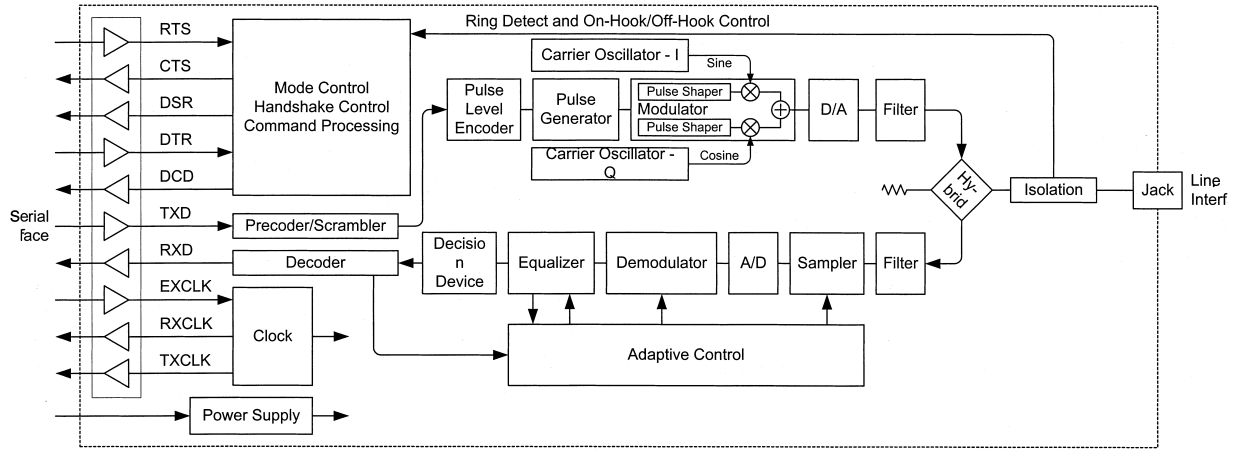


FIGURE 17 Full-duplex, two-wire, asynchronous, digital modem architecture.

The International Telecommunications Union—Telecommunications Standardization Sector (ITU-T) has issued a series of recommendations for modems at various speeds and for different applications (the V-series recommendations). All recommendations provide interworking between units made by different manufacturers. A number of the recommendations, as well as early AT&T Bell System standards used in the United States, are shown in Table V. The V.32 recommendation, which uses ITU V.29 automatic adaptive channel equalization, was the first to specify trellis coded modulation to improve the SNR. This modem type also uses echo cancellation. The combination of TCM and echo cancellation made dial-up operation possible with full-duplex data rates of 4.8 to 14.4 kb/s. The later ITU V.34 recommendation allows for speeds up to 33.6 kb/s.

ITU-T recommendations V.90 and V.92 are unique compared to previous recommendations in that they include both analog and digital operating modes and asymmetric rates. Recommendation V.90 incorporates the requirements of the V.34 recommendation for the upstream direction (subscriber to central office) with speeds up to 33.6 kb/s. Recommendation V.92 increases the maximum upstream speed to 48 kb/s. In the downstream direction, central office to subscriber, the V.90 and V.92 modems take advantage of an all-digital switching and transmission facility (if one exists) from the data network, such as an internet service provider, to the end office switching system. If only a single digital-to-analog conversion takes place between the end office and the subscriber's modem (typically in the end office line circuit), a line speed of 56 kb/s is theoretically possible. If more than one A/D conversion takes place between the data network and the subscriber or if the line quality will not support the higher speeds, the V.90 and V.92 modems drop back to the analog mode V.34 and downstream speed is limited to a maximum of 33.6 kb/s. In this configuration, the modem operates in a symmetric mode. The 48 kb/s speed of the V.92 modem in the upstream direction is possible by using a signaling rate of 8000 symbols/s and synchronizing the signals with the central office CODEC 8-kHz sampling clock. This effectively reduces problems due to modem clock recovery and errors associated with incorrect sampling of the received signal.

Many different modems have been developed over the years, each having higher speed and more complexity than its predecessor. All modems used in the public network have common user and network interfaces. Typical interfaces on the user side of the modem are IEA-232, ITU V.35, and Universal Serial Bus (USB) for external modems and PCI and ISA bus interfaces for internal PC modems. These interfaces exchange unipolar (unbalanced) or bipolar (balanced) baseband signals between the users data terminal equipment (DTE, such as a PC or router) and the data communications equipment (DCE, or modem). In the United States, the modem circuitry connected directly to the network must be electrically isolated from other modem circuitry to prevent a fault in the modem from causing network harm as required by FCC Part 68 [47 Code of Federal Regulations, Part 68]. This is commonly done through circuitry that provides isolation a well as supervisory (on-hook and off-hook) and ring detection functions. Isolation traditionally has been through coupling transformers and electromechanical relays but all solid-state versions are in common use.

Another characteristic of modern modems is the extensive training procedures a modem follows when it connects to another modem at the other end of the connection. These procedures include the initial tones used by the calling and answering modems, the length of the tones, and the types of modulation for the exchange of training information. For example, ITU V.34, V.90, and V.92 modems exchange information (INFO) sequences with binary DPSK modulation at 600 b/s using a 2400-Hz carrier plus a 1800-Hz guard tone. The INFO sequences include line probing results (to set equalizers), cyclic redundancy check (CRC) bits for error detection, frequency offset as measured at the receiver, power reduction to be implemented by the calling transmitter, pre-emphasis filter settings, and symbol rate to be used in transmitting from the answer to calling modem. In addition to the INFO sequences, a number of modulation parameters (MP) are exchanged during training to define the setting for trellis encoders, data signaling rates, and constellation shaping. In the V.90 modem, the complete training sequence takes around 20 to 30 s. This is reduced to 10 to 15 s in the V.92 modem.

As previously mentioned, low-speed modems are used for transmission of Caller Number and Caller Name information from the serving central office to the subscriber. This is a one-way simplex transmission from the central office to the subscriber and in this way differs from most other modem applications, which operate in a full-duplex mode. The transmission uses FSK modulation at 1200 b/s with carriers at 1200 Hz (logic 1 or space) and 2200 Hz (logic 0 or mark). The low bandwidth requirements of this low-speed modulation result in a robust transmission that is affected only by the worst subscriber lines. The information is sent as asynchronous (using start and stop bits), 8-bit ASCII characters between the first and second ringing signal on the loop.

IV. BASEBAND SIGNALING

Baseband (or unmodulated) signaling uses frequencies approaching zero (dc). There are two types of baseband

TABLE V Modem Recommendations and Standards

Recommendation or standard	Line rate	Half- or full-duplex	Transmit–receive separation	Carrier frequency (Hz)	Modulation	Symbol rate (Baud)	Bits per symbol	Synchronous or asynchronous	PSTN (2-wire)	Private line 2-wire or 4-wire	Equalization	Scrambling
ITU-T												
V.21	300	Full	Frequency	1080 & 1750	FSK	300	1	Either	Yes	No	Not specified	Yes
V.22	600	Full	Frequency	1200 & 2400	PSK	600	1	Either	Yes	2-wire point-to-point	Fixed	Yes
V.22	1,200	Full	Frequency	1200 & 2400	PSK	600	2	Either	Yes	2-wire point-to-point	Fixed	Yes
V.22bis	2,400	Full	Frequency	1200 & 2400	QAM	600	4	Either	Yes	2-wire point-to-point	Fixed/adaptive	Yes
V.23	600	Half	N/A	1300 & 1700	FM	600	N/A	Either	Yes	No	Not specified	Not specified
V.23	1,200	Half	N/A	1300 & 1700	FM	1200	N/A	Either	Yes	No	Not specified	Not specified
V.26	2,400	Full	N/A	1800	PSK	1200	2	Synchronous	No	4-wire point-to-point or mulipoint	Not specified	Not specified
V.26bis	1,200	Half	N/A	1800	PSK	1200	1	Synchronous	Yes	No	Fixed	Not specified
V.26bis	2,400	Half	N/A	1800	PSK	1200	2	Synchronous	Yes	No	Fixed	Not specified
V.26ter	1,200	Either	Echo cancellation	1800	PSK	1200	1	Either	Yes	2-wire point-to-point	Fixed or adaptive	Yes
V.26ter	2,400	Either	Echo cancellation	1800	PSK	1200	2	Either	Yes	2-wire point-to-point	Fixed or adaptive	Yes
V.27	4,800	Either	Physical	1800	PSK	1600	3	Synchronous	No	4-wire	Manual	Yes
V.27bis	2,400	Half/Full	2-wire/4-wire	1800	PSK	1200	2	Synchronous	No	2-wire/4-wire	Adaptive	Yes
V.27bis	4,800	Half/Full	2-wire/4-wire	1800	PSK	1600	3	Synchronous	No	2-wire/4-wire	Adaptive	Yes
V.27ter	2,400	Half	None	1800	PSK	1200	2	Synchronous	Yes	No	Adaptive	Yes
V.27ter	4,800	Half	None	1800	PSK	1200	3	Synchronous	Yes	No	Adaptive	Yes
V.29	4,800	Either	4-wire	1700	PSK	2400	2	Synchronous	No	4-wire point-to-point	Adaptive	Yes
V.29	7,200	Either	4-wire	1700	PSK	2400	3	Synchronous	No	4-wire point-to-point	Adaptive	Yes
V.29	9,600	Either	4-wire	1700	QAM	2400	4	Synchronous	No	4-wire point-to-point	Adaptive	Yes
V.32	4,800	Full	Echo cancellation	1800	QAM	2400	2	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
V.32	9,600	Full	Echo cancellation	1800	TC QAM or QAM	2400	4	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
V.32bis	1,4400	Full	Echo cancellation	1800	TC QAM	2400	6	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
V.33	1,4400	Full	Physical	1800	TC QAM	2400	7	Synchronous	No	4-wire	Adaptive	Yes

TABLE V (*Continued*)

Recommendation or standard	Line rate	Half- or full-duplex	Transmit–receive separation	Carrier frequency (Hz)	Modulation	Symbol rate (Baud)	Bits per symbol	Synchronous or asynchronous	PSTN (2-wire)	Private line 2-wire or 4-wire	Equalization	Scrambling
V.34	2,400 to 33,600	Half/Full	Echo cancellation	1600 to 1959	TC QAM	2400–3429	2 to 10	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
V.90												
Downstream	56,000	Full	Echo cancellation	Not applicable	PAM	8000	8 max	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
Upstream	33,600			1800 to 1959	TC QAM	2400–3429	10 max					
V.92												
Downstream	56,000	Full	Echo cancellation	Not applicable	PAM	8000	8 max	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
Upstream	48,000				PAM	8000	8 max					
Bell												
103/113	300	Full	Frequency	2225 & 1270 (orig.) 2025 & 1070 (answ)	FSK	300	1	Either	Yes	No	Fixed	No
108	300	Full	Frequency	2225 & 1270 (orig) 2025 & 1070 (answ)	FSK	300	1	Either	Yes	No	Fixed	No
201	2,400	Half	None	1800	PSK	1200	2	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
202	1,200	Half	None	1200 & 2200	FSK	1200	1	Either	Yes	2-wire point-to-point	Fixed	No
208	4,800	Half	None	1800	QAM	1600	3	Synchronous	Yes	2-wire point-to-point	Adaptive	Yes
212	1,200	Full	Frequency	1200 & 2400	PSK	600	2	Either	Yes	No	Fixed	Yes

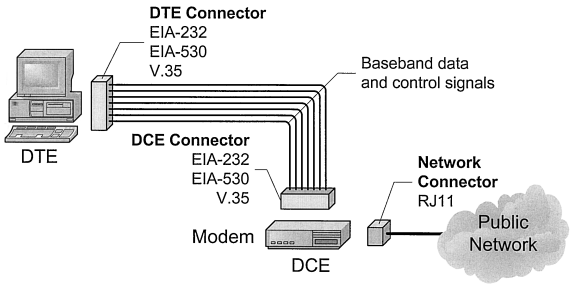


FIGURE 18 Data terminal equipment and data communication equipment.

signaling that apply to data transmission. The first are the baseband signals used between data terminal equipment (DTE) and data communication equipment (DCE), such as a modem (Fig. 18). Typically, unipolar or bipolar pulses are used to convey data values and control signals between DTE and the DCE. The types of signals used for these interfaces, which are defined in various industry standards, are suitable for short distances only. The modulation functions in a modem transform the baseband unipolar and bipolar signals into a passband line signal suitable for transmission through bandlimited analog channels on the network side of the modem.

The second type of baseband signals are those coupled directly (without a modem) to subscriber loops. In this case, an interface device called a Data Service Unit (DSU) or Channel Service Unit (CSU) is used instead of a modem. These devices reformat or convert the DTE/DCE interface signals so they are suitable for direct transmission on subscriber loop cables. The DTE/DCE interfaces of the DSU or CSU are the same as discussed earlier for a modem The network interfaces, on the other hand, use baseband digital pulses whose spectral characteristics are fundamentally defined by the pulse interval, amplitude, and width. In general, unipolar pulses are not used on loops because it is necessary to maintain a balanced condition on twisted cable pairs for electrical protection and crosstalk reasons. Instead, balanced bipolar pulses are used. Balanced pulses eliminate the dc component so that coupling transformers can be used for isolation purposes.

The bandwidth required by baseband signaling is directly related to the data speed by the Nyquist bandwidth constraint: A system with bandwidth $W = 1/2T$ Hz can support, without intersymbol interference, a maximum transmission rate of

$$R_s = 2W, \tag{2}$$

where R_s is the transmission rate in samples/second and T is the symbol interval in seconds.

A voiceband channel with bandwidth limited to 3.1 kHz will support a baseband signaling speed of 6200 symbols/s with no ISI. Each symbol can represent M levels of k bits, where $M = 2^k$. For example, when k = 5 bits/symbol, each symbol represents one of 32 levels, and the voiceband channel theoretically will support 31,000 b/s. This combination results in a spectral efficiency of 10 b/s per Hz. From Eq. (1), the required SNR = 30 dB. However, due to filtering and distortion introduced into the channel by A/D converters and the transmission medium, this speed is attainable only with a great amount of signal processing.

The downstream direction of the V.90 and V.92 digital modems use pulse amplitude modulated (PAM) baseband signaling with a sample rate, R_s, of 8000 samples/s. A PAM signal has amplitudes (voltage levels) corresponding to the bit values being transmitted during the symbol interval. The 8000 samples/s rate and 8 bits/sample correspond to the sampling rate and quantization used in telephony CODECs. With 8 bits/sample, the modems can use $M = 256$ levels maximum (Fig. 20). From Eq. (2), the capacity limit $2W$ samples/s $= 16W$ bit/s, since each sample contains 8 bits. Therefore, the maximum channel capacity in a 3- to 4-kHz bandwidth is 48 to 64 kb/s. Due to present network limitations (such as the use of robbed-bit signaling) and regulatory limitations, not all of the 256 levels are usable in every sample and the higher speed is not achievable. The maximum advertised line speed for V.90 and V.92 modems is 56 kb/s, but 50 kb/s is a more practical maximum.

These high rates are possible only if there are no A/D converters between the digital modem and the subscriber's modem (remember that A/D converters add distortion, which increases noise in the decoded signal and, thus, lowers the SNR). With this constraint met, only a D/A converter would be located between the digital modem and the subscriber's line circuit as shown in Fig. 19. Note that Fig. 19 clearly differentiates between the digital and analog modems (if there is at least one A/D converter, then the modem falls back to V.34 analog mode on both ends). The D/A converter in the line circuit introduces no additional distortion to the signal toward the analog modem.

In going to more levels, the data rate of a PAM baseband line signal is increased with no change in the signal bandwidth, thus increasing spectral efficiency. The maximum factor increase in data rate (or, equivalently, spectral efficiency) is $\eta = 2\log_2(M)$. However, due to filter rolloff, a more practical value is $\eta = [\log_2(M)]/(1+\rho)$, where $\rho =$ filter rolloff factor (defined in the following).

These types of modems require signal processing techniques involving a number of tradeoffs

- Desirable spectral characteristics can be realized through shaping the frequency spectrum of the pulse or the frequency spectrum of the correlated amplitude sequence, increasing computational complexity.

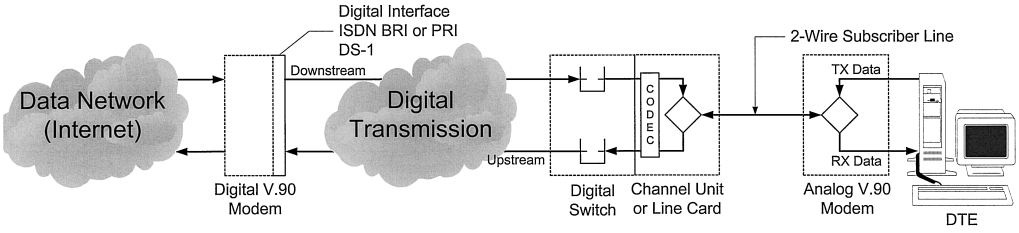


FIGURE 19 ITU V.90 digital and analog modems.

- Bandwidth can be reduced at the expense of noise immunity, a possibility implied by the Shannon–Hartley capacity formula [Eq. (1)]. For the same average power, using more signal levels reduces immunity to noise (there is less space between signal elements making it more difficult to decode the correct signal immersed in noise).
- Although pulses may overlap, they can be designed to not interfere with each other or to interfere in controlled ways. Intersymbol interference caused by linear channel distortion can be largely compensated for by further pulse shaping in both the transmitter and receiver.

A rectangular pulse—or any pulse wholly contained within one symbol interval of T seconds—does not interfere with any other pulse. A pulse constrained to a T-second interval, however, consumes more bandwidth than is necessary. The only requirement a pulse has to satisfy, to avoid intersymbol interference, is that all pulse samples at intervals T be zero except the one that carries the information. The Nyquist pulse (Fig. 21a) satisfies this condition with the minimum possible bandwidth of $1/2T$ Hz (a normalized frequency of 1.0 in the filter transfer function response of Fig. 21b). Since it is difficult to generate and work with pulses that have such an abrupt frequency spectrum (roll-off factor $r = 0.0$), actual modems will use a pulse with slightly greater bandwidth and smooth frequency characteristics while retaining the property of zero samples at symbol intervals. The frequency characteristic of a raised cosine pulse (Fig. 19a) is often used. The edge of the frequency characteristic is a cosine function rolling off in an additional fraction of the minimum $1/2T$ Hz bandwidth (in addition to the Nyquist pulse with $r = 0.0$, roll-off factors of 0.15, 0.5, and 1.0 are shown in Fig. 21). The selection of a filter (and, thus, the pulse shape) depends on the relative noise and bandwidth constraints in a given application. Raised cosine pulses with rolloffs of 0.1 to 0.33 are often used.

V. CHANNEL EQUALIZATION

The analog voiceband channel acts as a lowpass filter that spreads and smears the original, discrete time sequence of transmitted signals into a continuous time signal at the

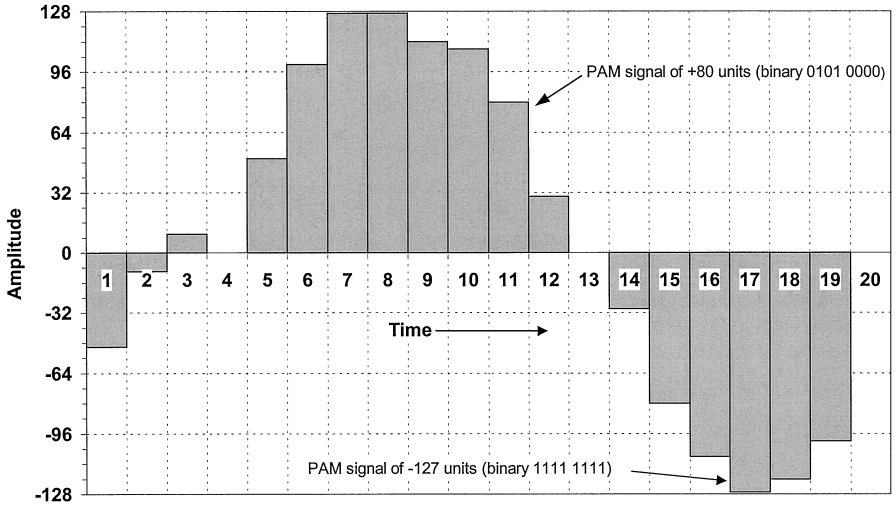


FIGURE 20 Pulse amplitude modulated signal. This example shows a PAM signal with maximum amplitude range of 256 units (−127 to 0 and 0 to +127), each corresponding to an 8-bit value. The most significant bit indicates the sign (+ = 0, − = 1).

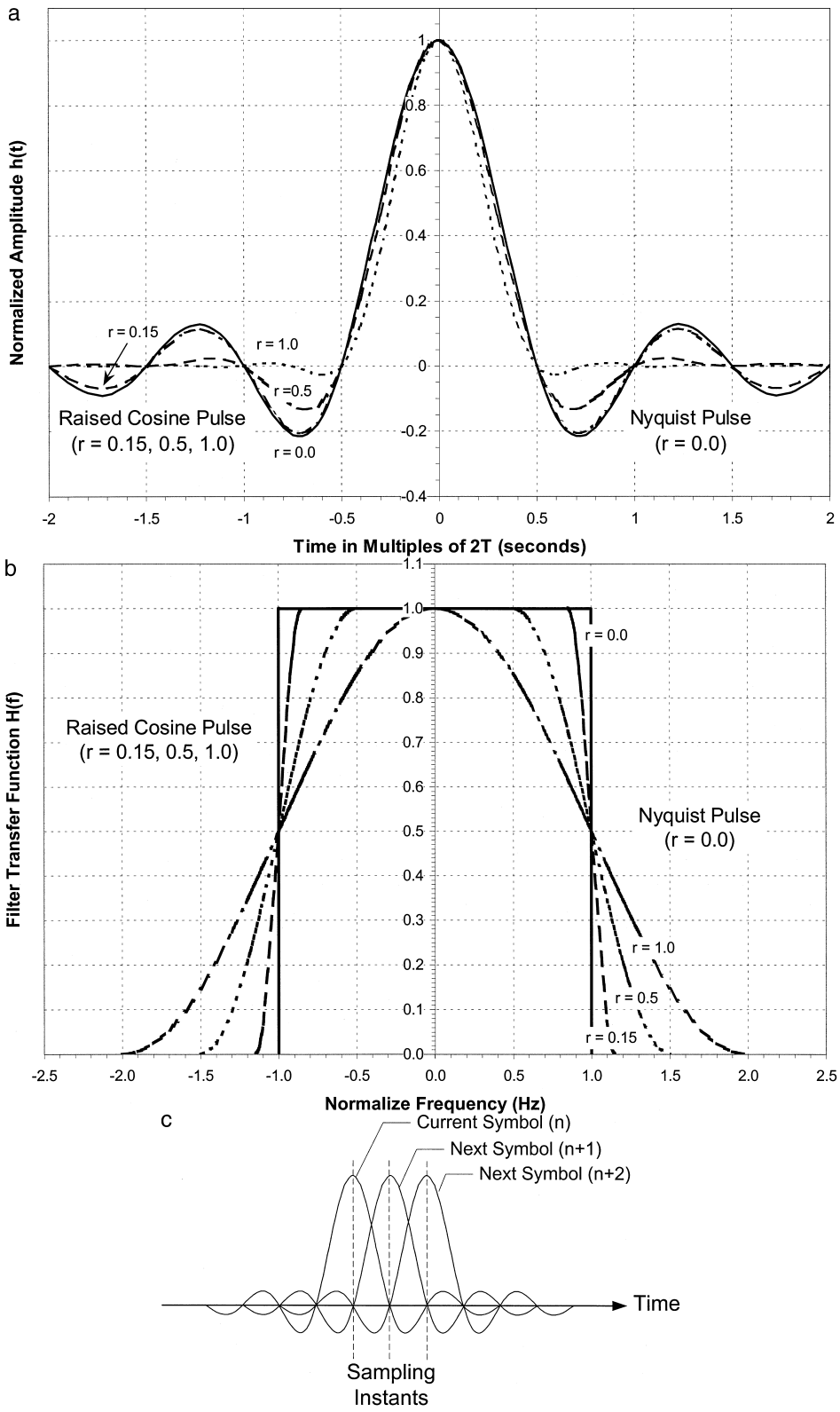


FIGURE 21 Band-limited pulses that do not contribute intersymbol interference. (a) Nyquist pulse and raised cosine pulses. (b) Frequency characteristics of pulses shown in (a). (c) Consecutive raised cosine pulses do not cause intersymbol interference because they have zero amplitude at the sampling instant.

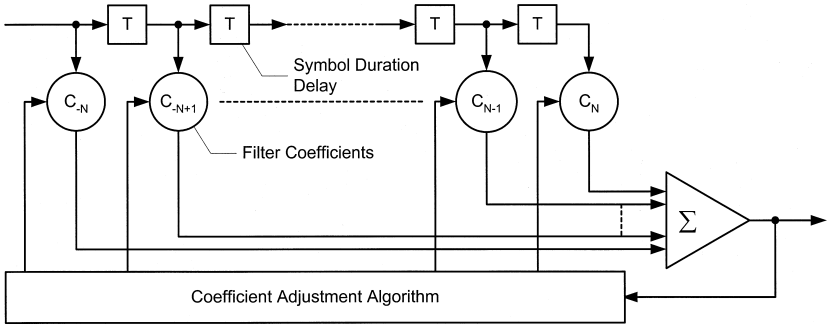


FIGURE 22 Transversal filter with 2N + 1 taps. The output of the taps are amplified, summed, and fed to a decision device that adjusts the tap coefficients to subtract the effects of interfering symbols that are adjacent in time to the desired symbol.

receiver. The channel also is corrupted by noise. The peaks of the received signal relate to the amplitudes of the original pulses.

If the channel characteristics are exactly known, a receiver can be designed that decides the transmitted signal sequence most likely to have produced the observed received signal. Such a receiver, called a maximum likelihood sequence estimation (MLSE) receiver, is difficult to implement directly, since the number of computations increases exponentially with the length of the sequence and the sequence could contain millions of symbols. In practical systems, the receiver equalizer filter contains parameters that are adjusted according to measurements of the channel. An equalizer can dramatically improve the performance of a modem over a communication channel, compensating for both amplitude and delay distortion in the channel.

A transversal filter, shown in Fig. 22, is a common choice for this job. The transversal filter is a delay line with taps spaced at intervals of the symbol duration (T). The outputs of the taps are adjusted via tap coefficients, c_n, summed and fed to a decision device. Signal distortion is corrected by selecting tap coefficients to force the output to zero at N sample points on either side of the desired pulse. In this way the effects of intersymbol interference are subtracted out.

The optimum receiver is a matched filter followed by a transversal filter, also called a Fractionally Spaced Equalizer (FSE). A matched filter is a linear filter designed to provide the maximum SNR at its output for a given transmitted symbol waveform. The impulse response of a matched filter is the mirror image of the input message signal delayed by the symbol time duration, T. Early implementations of this equalizer typically had 32 taps; modern high-speed modems have many more. The values of the tap coefficients are selected to reduce intersymbol interference through some performance criterion. The equalizer shapes the overall system impulse response so all symbol-interval samples are very small but the one of interest.

Another type of equalizer, useful when postcursors (the nonzero symbol-interval samples following the desired information-bearing sample) are significant, is the decision-feedback equalizer (DFE, Fig. 23). The DFE consists of a feed-forward filter and a feedback filter. The design presumption, as in the other equalizers, is that the channel impulse response is accurately estimated, so the postcursor values can be used as tap weights in the DFE to cancel the postcursor intersymbol interference entering the subtractor. Also, the signal improvement depends on the assumption that the decision data levels fed back through the delay line are almost always correct. When a decision error is made, performance can degrade for more than one symbol interval through error propagation.

Error propagation in a DFE is reduced by precoding. The ITU recommendation V.34 modem splits the DFE operation between the transmitter on one end and the receiver on the other end. The answering modem receiver calculates the optimum equalizer coefficients during training

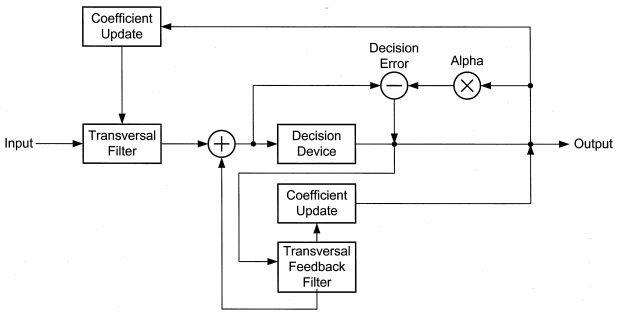


FIGURE 23 Decision-feedback equalizer. The output of the decision device feeds a tranversal filter that compensates the decision device. The second transversal filter in the feedback path cancels intersymbol interference on future symbols due to the current symbol.

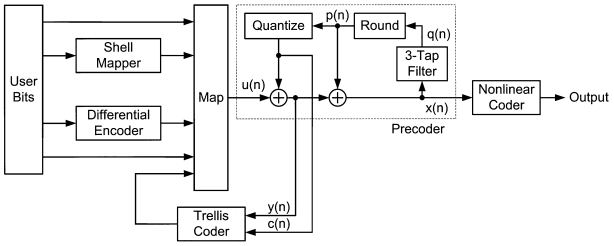


FIGURE 24 V.34 modem transmitter with precoder to reduce error propagation associated with a decision feedback equalizer.

and sends them to the calling modem. The calling modem transmitter then precodes the signal before transmission using a simple three-tap precoding filter (Fig. 24), representing the feedback filter portion of the DFE. The precoder "dithers" the transmitted sequence in such a way that the received signal sequence is free of linear distortion and only contains the trellis-coded sequence plus noise. This allows the receiver to decode the input signals by a simple feedback-free operation, thus limiting error propagation.

All equalizers depend on knowledge of the voiceband channel impulse response characteristics to adjust tap coefficients but, in practice, the channel characteristics are not precisely known nor do channel characteristics always remain static after the training sequences. Effective channel equalization calls for an adaptive equalizer that automatically adjusts tap coefficients to the best settings and thereafter tracks any changes in the channel. The tap coefficients of the transversal equalizer can be driven close to their optimum values through the classical Minimum Mean Square Error (MMSE) adaptation algorithm.

The initial training of an equalizer does not work well if the starting (unequalized) error rate is high. To get around this, a known data sequence is often used at first. Very rapid training is possible with a special short data sequence equal in length to the number of symbol intervals spanned by the equalizer. This cyclic equalizer converges in the same number of symbol intervals. An alternative start-up technique is to transmit first at a lower speed, for which the error rate is not too high, and use data decisions for adaptation.

An additional start-up problem occurs in polling applications, where a central modem interrogates a collection of remote modems. A significant adaptation delay each time the host turns to a different remote device is highly undesirable. Time can be saved by the central modem remembering its tap weight settings from its last contact with the previous remote device and going through an abbreviated recovery process, using a special data sequence, to recover timing and other parameters consistent with the saved tap weights.

VI. PASSBAND DATA COMMUNICATIONS

Passband data communications involves sending data through a telephone channel with passband characteristics similar to Fig. 10. The data are in the form of a baseband information signal that is used to modulate a carrier waveform. The telephone channel is narrowband compared to the information rates that users want to send through it, so spectrally efficient modulation formats are used. Linear modulation techniques, which code data streams into two PAM pulse trains impressed on sine and cosine carriers (that is, carriers with 90° phase difference), respectively, achieve a frequency translation from baseband to passband that does not expand the bandwidth. The spectral efficiency of multilevel PAM carries through to these "two-dimensional" modulated signals, as do the structures of channel equalizers and much of the analysis of their performance. The main additional feature is adaptive phase tracking circuitry that must function simultaneously with the adaptive channel equalizer.

Before describing the structure of a 2-dimensional modem, which was introduced in Fig. 17, it is worth noting that at lower data rates, for which spectral efficiency is not so critical, frequency shift keying (FSK) techniques work well. FSK modems operating full duplex at 1200 b/s on dial-up lines were common in the mid-1980s. In this type of modem, the analog channel was separated into two halves, using a separate carrier for each direction of transmission as shown in Fig. 8a. In multilevel FSK, each level uses a different carrier frequency, with the deviation ratio selected to achieve a desirable spectral shape (for example, one without spikes). Unlike linear modulation, the FSK bandwidth increases with the number of levels. FSK has the advantage of a constant signal amplitude (envelope), which makes it robust against non-linearities of power-limiting devices in the channel. If care is taken to make it phase continuous at the boundaries between symbols, the spectrum for binary FSK is only slightly wider than that for binary amplitude modulation.

Phase modulation also produces a constant amplitude passband signal but is a linear format like amplitude modulation. Figure 25 shows the general forms of a modulator and demodulator for M-phase (M-ary), phase-shift keying (PSK) and the two-dimensional signal constellations for two- (binary), four-, and eight-phase PSK. A two-dimensional signal constellation is the set of possible amplitude pairs, or signal points, for the sine and cosine carrier waveforms or, equivalently, the set of magnitude and phase pairs. The constant-magnitude signal points of PSK shown in Fig. 25 correspond to 2-, 4- and 8-phase modulation (1, 2 and 3 bits per symbol, respectively).

The bit error probability increases rapidly with the number of phases above four and is especially vulnerable to

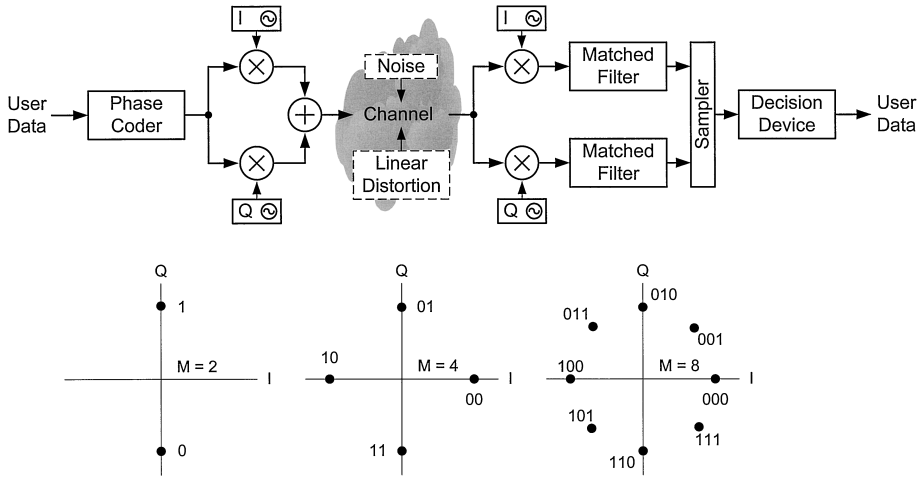


FIGURE 25 PSK transmitter and receiver and signal constellations for two-, four-, and eight-phase systems with rectangular pulse shapes.

phase jitter. In order to avoid the carrier recovery problem (local oscillator reference phase determination), differential PSK (DPSK) may be used. In this case, data bits are coded into a change in phase rather than an absolute phase, but at a price in error rate performance for a given SNR.

Phase modulation is a two-dimensional signal format with the signal points constrained to lie in a circle. Higher spectral efficiency and better error performance can be attained with 16 or more signal points distributed more uniformly in the two-dimensional space rather than in a circle. A more uniform distribution increases the minimum distance between signal points and thus the immunity to error caused by noise, phase jitter, and other impairments. A more uniform distribution is provided by quadrature amplitude modulation (QAM), which is a combination of phase shift keying (PSK) and amplitude shift keying (ASK). QAM uses two independent information channels to be modulated onto a single carrier. These two channels are called the real (or In-phase) and the imaginary (or Quadrature) components of the signal. Each point on the resulting constellation has a unique set of real and imaginary components and is used to represent a group of bits. Two 16-point QAM constellations (with 4 bits representing each point) are compared in Fig. 26.

When plotting the error performance of different types of modem signaling, it is convenient to use E_b/N_0 rather than SNR, where E_b is the signal energy per binary symbol, or bit, and N_0 is the noise in a bandwidth of 1 Hz. This allows different signaling methods to be easily compared side-by-side on the same plot. Figure 27 shows bit error probability versus E_b/N_0 for M-ary PSK with $M = 1$ and 3 bits/symbol (2- and 8-point, respectively) and for 16- and 64-point QAM.

The transmitter for data communications with two-dimensional signals is essentially the same as that in Fig. 25 with the addition of pulse-shaping filters. The receiver structure, including a passband equalizer, is shown in Fig. 28. This is an expansion of the structure shown in Fig. 17. It can be shown that it is equivalent to have the equalizer either precede or follow the demodulator. However, there is an advantage, for phase-tracking demodulation systems, to take the delay introduced by the equalizer out of that adjustment loop. Whether in passband or baseband format, the equalizer must be designed to process a complex analytical pair of signals differing in phase by 90° in each frequency component. Such an equalizer is shown in Fig. 29 following the demodulator for analytical convenience. The incoming signal is converted to a complex analytical signal by creating a new "imaginary" signal in a Hilbert filter, which is a 90° phase rotator.

Introducing redundancy to the transmitted signal stream doubles the constellation size. This is not a problem for small constellations but detecting constellation points due to increased density becomes a problem in larger constellations. With trellis coding, redundancy is provided in a

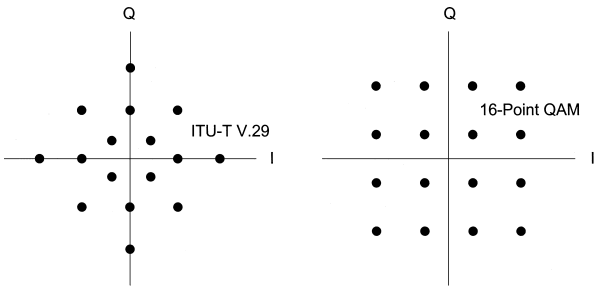


FIGURE 26 Two-dimensional, 16-point QAM signal constellations. (a) ITU-T V.29 QAM constellation. (b) Another example of 16-point QAM constellation.

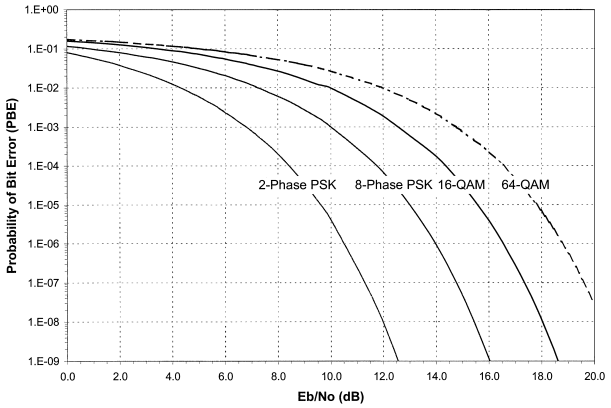


FIGURE 27 Bit error probabilities for two- and eight-phase PSK compared to 16- and 64-point QAM.

two-dimensional space (doubling the constellation size). It is possible to distribute the redundant bits in more than two dimensions in which case the number of bits per symbol mapped to a two-dimensional space is now less. For example, it is possible to achieve redundancy with 5 bits for every two symbols, or an effective $2\frac{1}{2}$ bits per symbol.

For constellations in rectangular two-dimensional space, the symbols at the corners require more energy than at the edges. This results in power inefficiencies in large constellations. A more power-efficient method is to shape the constellations in a circular pattern or, ideally, in a spherical pattern. One way of shaping the constellation is to project the sphere onto a two-dimensional plane for transmission in a single symbol; however, this results in a nonuniform circular symbol distribution with the highest symbol density at the center. This reintroduces the problem that distances between symbols are small when they are close to the center. Shell mapping helps to overcome this problem. Shell mapping algorithms used in V.34 modems use 16 dimensions with constellations mapped across 8 QAM symbols (each QAM symbol has two dimensions, In-phase and Quadrature-phase). The shell mapping algorithms achieve near spherical constellation shaping with small increases in constellation size. The 33.6 kb/s rate of V.34 uses nested quarter-constellations that are generated as 416-point subsets (Fig. 30) of a single 1664-point QAM super-constellation. The full super-constellation is obtained by the union of the four quarter-constellations obtained by rotating the quarter-constellation 0, 90, 180, and 270°. The V.34 28.8 kb/s rate is achieved with a super-constellation size of 960 points consisting of four subsets of 240 points each using the same method as for the 33.6 kb/s rate.

The V.34 recommendation uses various combinations of bits per symbol and symbol rates to achieve a given line speed. For example, the 28.8 kb/s rate at 3200 baud results in 9 bits/symbol. This requires a constellation size of 896 points. To implement shell mapping for this example, the signal constellation is first partitioned into 14 equal-sized concentric rings of 64 points each. The encoder collects incoming user bits in frames of 72 bits over 8 symbol periods. Of these, 28 user bits are used by the shell mapper (Fig. 24), which selects one of the 14 rings for each of the 8 QAM symbols in the frame. The rings are chosen so that the 16-dimensional signal point lies in a quasi-spherical region. The remaining 44 user bits, together with 4 coded bits generated by the trellis encoder, are then used 6 at a time to select 1 of the 64 transmitted signal points in each selected ring.

Three major innovations have been developed for high-speed voiceband modems since the Fractionally Spaced Equalizer—trellis coded modulation (TCM), echo cancellation, and multisymbol line coding. Trellis coded modulation is a form of forward error correction (FEC) that convolutionally encodes the transmitted signal points (the term "trellis" refers to the appearance of the path diagram that depicts coder state changes). A trellis coding algorithm adds redundancy by accepting m data bits as input and generating $m+1$ bits as output. TCM is useful in bandwidth-limited channels, such as voiceband channels, because it provides FEC without increasing the bandwidth requirements. TCM combines modulation/FEC encoding and demodulation/FEC decoding in an integral operation that effectively increases the minimum distance between the signals that are most likely to be confused in the decoder. Sequences of trellis coded symbols are restricted to certain valid patterns. These patterns are chosen to have a large minimum distance between them so that corrupted sequences can be corrected, thus reducing the error probability of the decoded signals.

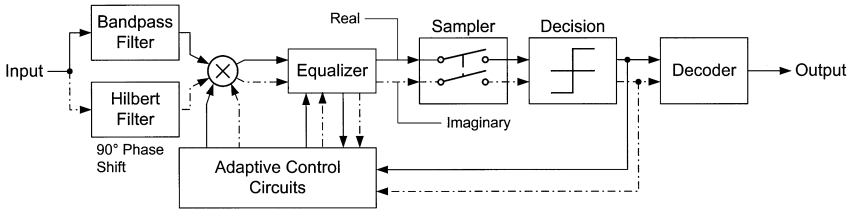


FIGURE 28 Receiver structure for two-dimensional signals. Complex variables simplify description and analysis.

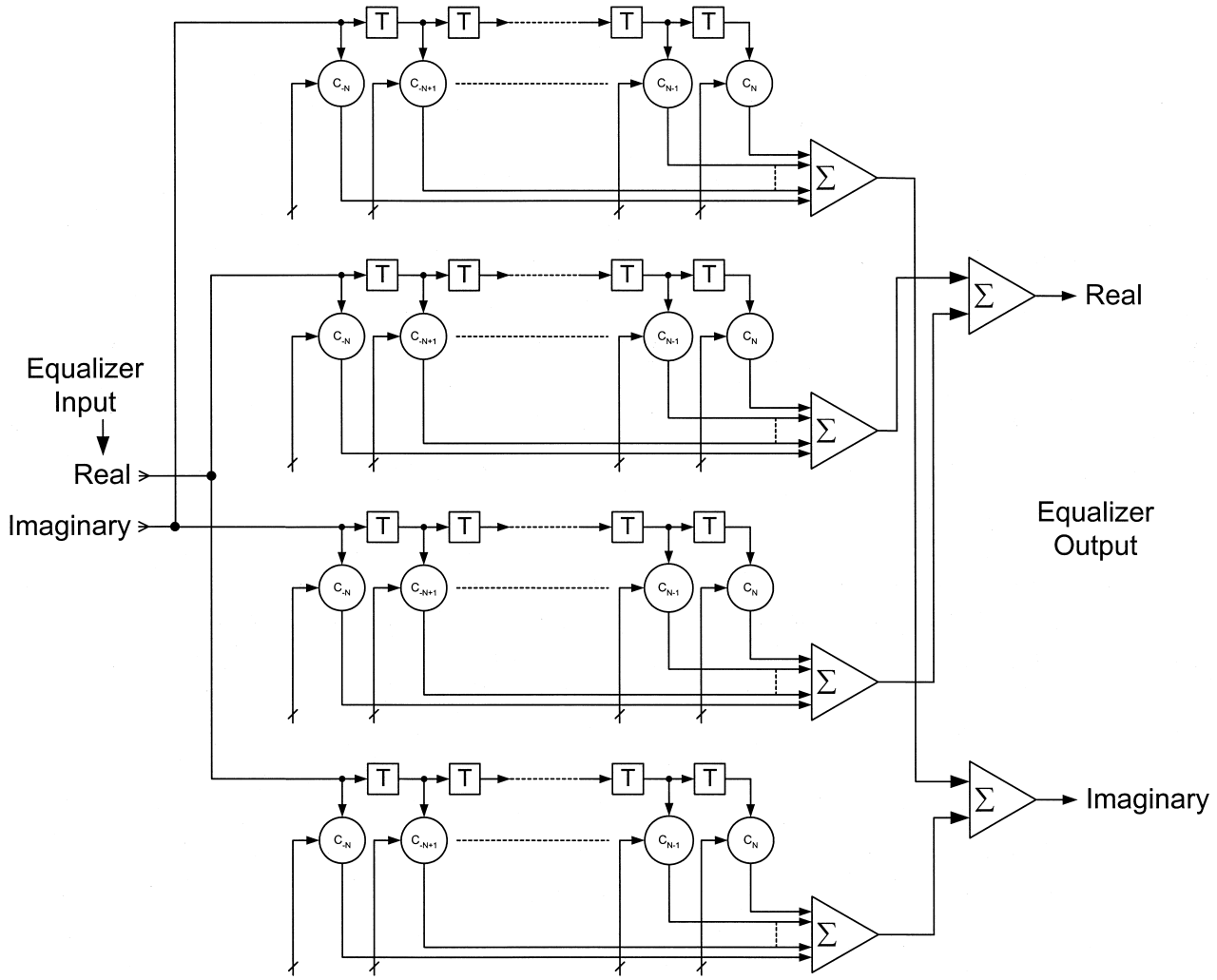


FIGURE 29 Cross-coupled equalizer structure of Fig. 28.

A Viterbi algorithm is used to decode the TCM symbols. A Viterbi algorithm is a maximum-likelihood process that discards low-probability paths and, thus, reduces the number of paths that must be considered during the decoding operation. The Viterbi decoder takes the output of the demodulator and equalizer, which is a sequence of noise-corrupted constellation points, and maps them to their ideal constellation points. The farther a received signal is from a point on the ideal constellation, the less likely it corresponds to that point. The Viterbi decoder relies on past history when deciding the value of the current point.

Figure 31 shows the convolutional encoder and the 32-point signal constellation used in the ITU recommendation V.32 modem (1990). The 32-point constellation with trellis coding is an alternative to the 16-point constellation used for 9600 b/s transmission without trellis coding. The partitioning procedure for assigning encoder outputs to signal points is an important element of trellis coding and is indicated by the bits-to-signal point encodings shown under the signal points in the illustration. An improvement in SNR of several dB is possible by using this technique on severely distorted channels.

In V.32, the 9.6 kb/s transmit data stream is divided into groups of four bits. The first two bits ($Q1_n$ and $Q2_n$) are differentially encoded and then convolutionally encoded to generate three new encoded bits ($Y0_n$, $Y1_n$, and $Y2_n$). The other two bits ($Q3_n$ and $Q4_n$) are not encoded but sent directly to the output stage, where they are mapped with the other bits into the signal space constellation in a predetermined way. The encoded bits, $Y0_n$, $Y1_n$, and $Y2_n$, divide the 32-point constellation into eight subsets of four modulated carrier states. Each combination of $Y0_n$, $Y1_n$, and $Y2_n$ identifies a subset. The unencoded bits, $Q3_n$ and $Q4_n$, identify a point within the subset. Note that, in Fig. 31, all points beginning with the same three bits ($Y0_n$, $Y1_n$ and $Y2_n$) are "widely" separated. This type of

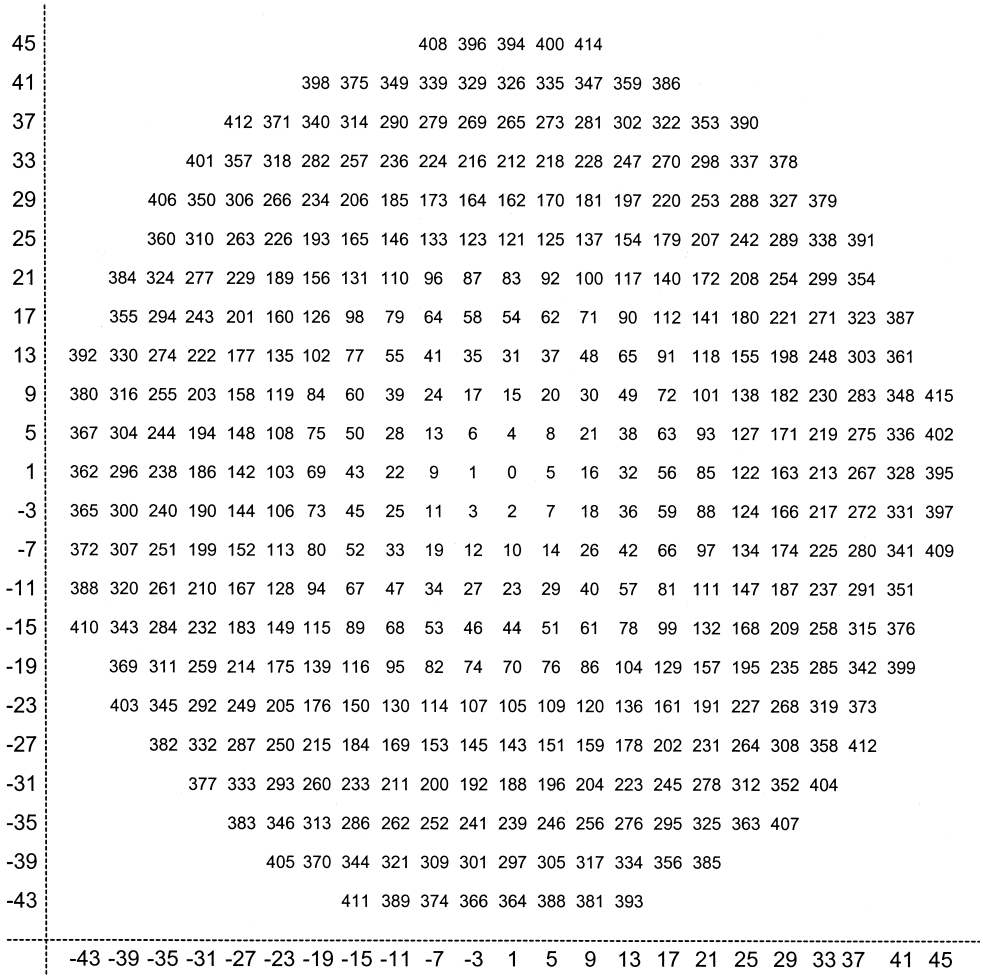


FIGURE 30 ITU V.34 quarter-constellation subset of 416 points.

partitioning and the presence of the redundant bit ($Y0_n$) form the basis for trellis coded modulation. The V.32bis recommendation (1991) for 12.0 and 14.4 kb/s uses a similar principle except the data stream is divided into groups of 5 and 6 data bits, respectively. The trellis encoder for 12.0 kb/s outputs a total of 6 bits, which are mapped into a 64-point constellation, and the 14.4 kb/s trellis encoder outputs 7 bits, which are mapped into a 128-point constellation. The V.34 recommendation (1998) uses a 16-state four-dimensional trellis code, which provides a coding gain of about 4.2 dB, as well as 32-state and 64-state four-dimensional trellis codes. Although the V.34 transmitter must support all three codes, the V.34 receiver only has to support any one.

Echo cancelers are required wherever it is necessary to separate the transmit and receive directions on two-wire facilities and are used on all high-speed voiceband modems as well as on digital subscriber line (xDSL) technologies. Echo cancellation was shown conceptually in Fig. 8b. A more detailed illustration is given Fig. 32, which shows an echo canceler at each end of the connection. The hybrid coupler, which performs the two-wire to four-wire conversion both in the modem and in the analog line circuit, does not keep all of the transmitted signal out of the receiver (near echo) nor does it block the long-distance echoes that return from the hybrid couplers located at the far end of the circuit (listener or far echo). The desired receive signal is quite low compared to the strong transmitted signal and the echo produced by the transmitted signal.

Adaptive transversal filters, very much like channel equalizers, are used to generate replicas of the near- and far-end echoes, which are subtracted from the actual echoes. The filters normally are driven by the data levels rather than the transmitted signal in order to reduce the computational complexity. The filter adapts to the channel; in higher speed modems, this adaptation takes place using a probing signal during the training sequence prior

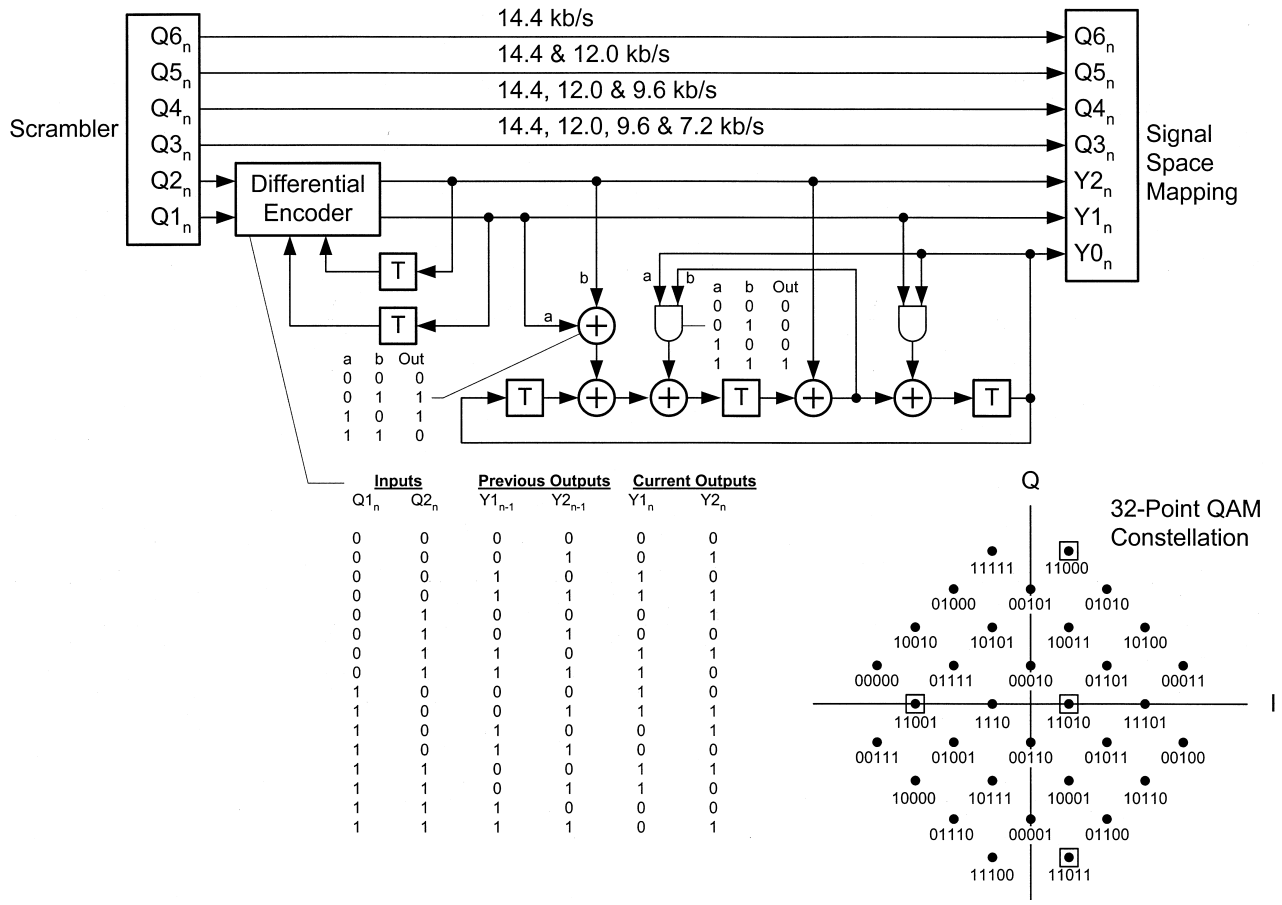


FIGURE 31 Convolutional encoder and signal constellation used in ITU recommendation V.32. The scrambled data stream is divided into groups of 3, 4, 5, or 6 consecutive data bits depending on the speed. The 32-point signal constellation shown is for 9600 b/s data communication. The binary number under each signal point is the data sequence coded into that point in the following order: $Y0_n$, $Y1_n$, $Y2_n$, $Q3_n$, $Q4_n$. The squares around the four signal points identify the signal points associated with Y0Y1Y2 = 110.

to actual data transmission. The bulk delay preceding the far-end echo canceler is determined from a sounding of the channel or some equivalent adaptation procedure. An echo canceler suppresses the echo by 60–65 dB and makes it possible to operate full-duplex at relatively high rates on dialup telephone lines.

The final innovation, multi-symbol line coding, codes an input data stream into pairs, triplets, or quads of successive symbols, thus increasing the minimum separation of signal points and the immunity to noise. A coding gain of 1–3 dB is possible. An example of this type of encoding is coding 8 bits into the 256 possible signal points in a four-dimensional signal space corresponding to two symbol intervals in a 16-point QAM system.

Another way to improve the performance of pass-band communications is to use pre-equalization (also called pre-emphasis) before transmitting each symbol. Pre-equalization pre-distorts the phase and amplitude of the carriers based on the anticipated or measured channel transfer function so the current symbol will be received correctly. This somewhat eases the amount of processing at the receiver.

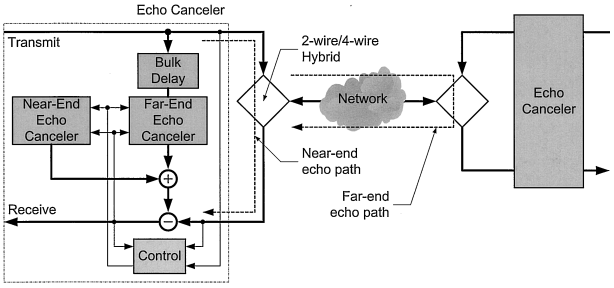


FIGURE 32 Echo canceler structure in a modem for full-duplex operation on two-wire dial-up telephone lines. The echo canceler subtracts a replica of the echo using two principal components—control mechanism and adaptive filters—to handle far-end and near-end echo.

VII. FUTURE OF DATA COMMUNICATIONS

Each new issue of ITU recommendations for analog voiceband modems is thought by some to be the last, but this is unlikely. As previously mentioned, ITU recommendation V.90 and V.92 modems use a combination format with a digital (baseband) format in the downstream direction (central office to user) and analog (passband) format in the upstream direction. This takes advantage of the current network topology and limitations. However, as the network evolves, users will move to an all digital format for many reasons, primarily the reduction in cost for higher bandwidth on digital subscriber lines. Digital equipment is more robust and easier to maintain, error performance does not significantly degrade with distance or number of tandem connections, different media can be used for transport, switching and processing, user interfaces are standard and the number of interfaces is comparatively small. After all, user bits are bits, and the user does not care how they are transmitted as long as they are transmitted quickly and received error-free.

Network operators (specifically, incumbent local exchange carriers) have made a huge investment in twisted pair plant and analog voice grade interfaces, and these will not be abandoned or replaced anytime soon. Voiceband data communications will continue to play a role in the public network as long as they are in place. However, users continue to demand higher speeds and lower costs, so xDSL technologies will play an increasingly important role. Such technologies presently are deployed in the public network and carry data at approximately 8 Mb/s over a twisted pair loop of around 3.6-km (2-mi) length. A version of xDSL, very-high-speed DSL (VDSL), can carry data on twisted pairs at speeds approaching 500 Mb/s over a short distance of 50 m and 100 Mb/s over a distance of 200 m. These types of DSL are not yet ready for widespread deployment in the subscriber loop. All twisted pair data transmission technologies use equalization, echo canceling and coding technologies originally developed for analog voiceband modems described in this article.

SEE ALSO THE FOLLOWING ARTICLES

BASEBAND DIGITAL LOOPS • COMMUNICATION SYSTEMS, CIVILIAN • DATA TRANSMISSION MEDIA • DIGITAL SPEECH PROCESSING • MICROWAVE COMMUNICATIONS • NETWORK FOR DATA COMMUNICATION • OPTICAL FIBER COMMUNICATIONS • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, DIGITAL • TELECOMMUNICATIONS • TELEPHONE SIGNALING SYSTEMS, TOUCH TONE

BIBLIOGRAPHY

TR-TSY-000333, Switched and Special Access Services, Transmission Parameter Limits and Interface Combinations, Telcordia.

ANSI T1.512-1994, "Network Performance—Point-to-Point Voice-Grade Special Access Network Voiceband Data Transmission Objectives," American National Standards Institute.

ANSI T1.506-1997, "Network Performance—Switched Exchange Access Network Transmission Specifications," American National Standards Institute.

IEEE Std. 743-1995, "Equipment Requirements and Measurement Techniques for Analog Transmission Parameters for Telecommunications," Institute of Electrical and Electronics Engineers.

47 Code of Federal Regulations, Part 68, Connection of Telephone Equipment to the Public Network, Federal Communications Commission.



VSAT (Very Small Aperture Terminal) Satellite Earth Station

Jonathon Y. C. Cheah
Uniden Corporation

I. Background
II. Basic VSAT Network Design Concept
III. VSAT Hub Station Design
IV. VSAT Station Design
V. VSAT Space Segment
VI. VSAT Applications
VII. Weaknesses of VSAT Networks
VIII. Other Specialized VSAT Networks

GLOSSARY

Downlink Transmission from the satellite to the earth station.
Equivalent isotropic radiated power The total output power of a transmitter multiplies with the gain of the antenna used.
Feed horn A device used in a reflector antenna so that the RF energy can be launched into the air.
Forward error correction A means of correcting errors in the received digital data.
Geosynchronous Describes a pseudostationary position in the satellite orbit above the equator whereby a satellite in such a position would appear to be stationary with respect to an observer on the ground.
Isochronous Describes the on-time delivery of data in a discontinuous packet form.
Noise figure A measure of the ambient noise that exists in a system.
Power flux density A generic measure of the radiated power by dividing the total signal power received by its bandwidth to give a per-hertz value.
Spectrum analyzer An instrument that displays the spectrum contents of a signal in terms of amplitude versus frequency.
Uplink Transmission from the earth station to the satellite.

VSAT is a physically small satellite earth station dedicated to digital data communications for private network use. Today, the worldwide VSAT population is estimated in the industry to be slightly less than a half million stations. A VSAT satellite network is a popular communications

option to provide medium speed-data traffic in point-to-multipoint network configurations. It is a particularly popular choice for private ownership of a digital communications network, which has geographically dispersed terminal connections. In many countries, VSAT technology is perhaps one of a very few ways whereby a private citizen or an organization can own and control all segments of a dedicated communications infrastructure without being firmly under the auspices of a government PTT (post, telephone, and telegraph) organization. VSAT technology further owes its popularity to its relatively economical data traffic tariff in terms of "per data bit transferred" cost in comparison to its terrestrial counterpart. Also, it has the advantage of the ease in deploying and commissioning a new communications network within a very short time.

I. BACKGROUND

The past 20 years has seen a great advancement in all areas of telecommunications technology, and without exception, digital satellite communications technology has also made giant strides during this period. In this case, the most noticeable outcome is the private ownership of small satellite networks. The reasons behind the need to own one's network for data-traffic-intensive users are many, nevertheless, the more pertinent catalyst was the impetus of change generated by the AT&T divestiture in the 1980s (Lyon, 1985, Chakraborty, 1988), and the uncertainty that created. There was a general paradigm shift among the heavy telecommunications users toward finding cheaper and better ways of connecting data terminals together, with the emphasis on the ability to administer all aspects of the communications operations within the network. Indeed, during the early 1980s, the few significant pioneering early adopters of the VSAT technology, each with over a few thousand VSAT stations in the network, expressed similar interest in network ownership requirements. Network ownership permits the quality of service and the grade of service to be completely manageable and predictable by the network owners. Of course, the one overriding requirement was that VSAT technology must provide the lowest cost in the "per data bit transferred" criterion against all other possible options at the time.

There were many difficult hurdles in realizing the promise of VSAT in the early days. All elements of satellite communications were and still are, in general, strictly regulated by international bodies such as International Telecommunications Union (ITU). Elaborate rules of frequency and site coordination procedures are applied. In addition, every host country has stringent regulations governing the use and operation of satellite earth stations. The fundamental regulatory technical requirements in regards to geostationary satellite earth station operations, however, are related to the intensity of the downlink satellite signal at ground level in terms of power spectral flux density, and that of the spatial spilover adjacent-satellite interference from the uplink signal of the earth station. There was another interesting problem. Many U.S. cities and towns have strict aesthetic ordinances governing the presence of large antenna structures on private properties, a remnant backlash arising from the excesses of the C-band television-receive-only (TVRO) industries in the United States during the late 1970s. Thus, the earliest known VSAT solution to overcome these hurdles was a C-band geostationary satellite earth station using a wideband spread-spectrum signal. In this way, it was thus possible to use a center-fed prime-focus circular antenna with a diameter of less than 1 m to satisfy these regulatory requirements. The majority of these early VSATs operated in the receive-only mode. This shortcoming limited the overall user acceptance of the C-band spread-spectrum VSAT solution.

In 1984, FCC promulgated the "VSAT order" in FCC Docket No. 81-704, and a subsequent Ku-band order (FCC 51, 1986) that adopted certain revisions to the original 47 CFR Part 25 of the FCC rules regarding satellite communications. In effect, the FCC "VSAT order" regulates the VSAT antenna sidelobe levels and the satellite downlink power flux density limits on the earth surface. It relaxes the stringent frequency coordination requirements and allows approval for blanket VSAT station licenses. The regulation works well in enssuring that acceptable satellite interference power flux density on the ground is maintained, and it protects the integrity of the 2° satellite parking spacing allocations in the geostationary satellite orbit. The power flux density provisions of the relevant FCC rules can be found in 47 CFR 25.134 and 47 CFR 25.208 and the antenna technical requirements are contained in 47 CFR 25.209. The VSAT industry was born on the strength of this FCC ruling.

II. BASIC VSAT NETWORK DESIGN CONCEPT

A small earth station is physically limited by a small antenna diameter. A small antenna poses difficult problems in establishing a good satellite link. The small antenna has low antenna gain in the transmit frequencies. The cost and regulatory elements limit the amount of transmitter power a station can possess. The small antenna also has low antenna gain in the receive frequencies, and again, cost constraints limits the ability of the low-noise amplifier (LNA) that can be used in the receiver. Thus, a small earth station has an intrinsically poor receive-noise figure

(NF), a high antenna sky-noise floor, and a low transmit equivalent-isotropic radiated power (EIRP) performance. From the satellite link budget analysis point of view between two satellite earth stations, it is clear that these performance deficiencies of a small earth station can be amply compensated by having a second, large earth station as a partner. The large station possesses ability to provide good transmit EIRP to match the poor receive noise figure of the small station, so long as it is within the satellite downlink power flux density constraints. In addition, the good receive noise figure of the large station can be used to offset the low EIRP of the small station. Thus, in expanding this two-station relationship to a network of multiple small earth stations such that the large earth station serves as a central hub station, a natural "star connectivity" network topology evolved. Of course, it is also possible to have two equally capable earth stations communicating with each other in a single-channel-percarrier (SCPC) basis, or multiple equally capable earth stations in a mesh network situation. From the preceding discussion, these variations do not maximize network infrastructure cost and data transport efficiency to the same extent as the "star connectivity" network in many respects and for obvious reasons. In general, the "star connectivity" network topology predominates in the VSAT industry.

Figure 1 shows a pictorial representation of a typical VSAT network in a star-connection configuration. In the continental United States (CONUS), the physical-hub earth-station antenna sizes used in a Ku-band satellite network range from 4 to 10 m in diameter. The Ku-band VSAT antenna sizes typically range from 1.2 to 1.8 m in diameter. In areas where there is marginal satellite footprint coverage, VSAT antennas with 2.4-m diameter are often necessary. The selection of antenna sizes is the result of the satellite link budget calculation. There is an interesting observation from the link budget analysis that is necessary to highlight. It can be said that the feasibility of a VSAT network in terms of satellite link calculation benefits a great deal from the average additional 5 dB of coding gains derived from the powerful modern forward-error-correction (FEC) coding techniques.

An actual commercial VSAT network link budget for the network outroute signal worksheet is shown in Table I. Outroute path is referred to as the signal path from the transmitter of the hub station to the receiver of the VSAT station. The inroute link budget, the signal path from the transmitter of the VSAT to the receiver of the hub station, is shown in Table II. A more in-depth link budget calculation example including independent uplink and down link rain fade considerations can be found in Cheah (1999).

A VSAT network uses well-defined frequency channels within a satellite transponder frequency assignment. The network uses a time-division-multiplex (TDM) technique for the outroute path where data packets are sent continuously by the hub station. There are no intentional breaks in the signal transmission. The inroute path

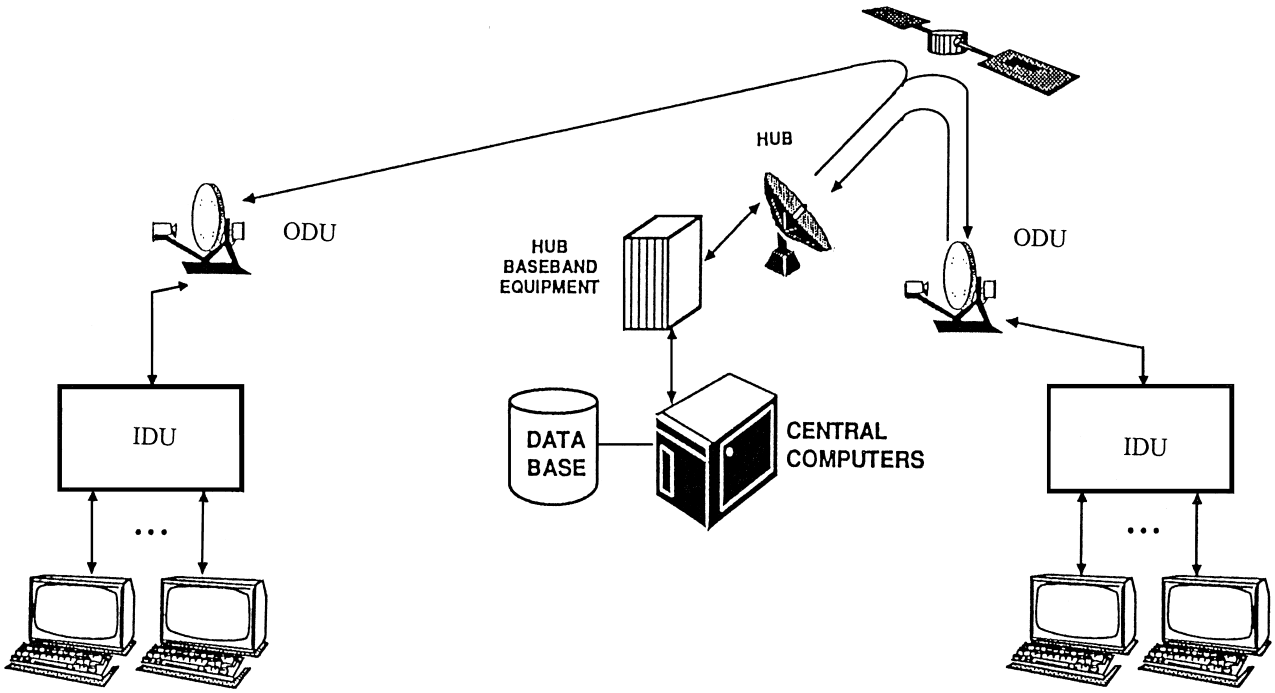


FIGURE 1 A pictorial representation of a typical VSAT network in a star-connection configuration. Each VSAT station is consisted of an outdoor unit (ODU) and an indoor unit (IDU).

TABLE I Outroute[a] Link Budget Calculation Example for a Commercial 800 VSAT Station Network Belonging to a Hotel Chain

Operating parameters					
Channel					
Modulation type				BPSK	
Transmission type				TDM	
Modem transmission rate				1024.0 ksps	
Code rate				$^1/_2$ sequential	
Threshold BER				1×10^{-7}	
Threshold Ebt/No				3.0 dB	
Transmit site					
Antenna diameter				7.0 m	
Frequency				14.250 GHz	
HPA size				110.0 W	
HPA operating power				1.3 W	
EIRP toward satellite				57.2 dBW	
Receive site					
Antenna diameter				1.8 m	
Frequency				11.950 Ghz	
LNA noise temperature				300.0 K	
Satellite					
Receive (G/T)				2.0 dB/K	
Saturation flux density				-87.0 dBW/m^2	
Saturated EIRP				40.0 dBW	
Transponder bandwidth				43 MHz	
Composite input backoff				10.0 dB	
Composite output backoff				6.0 dB	
Threshold carrier-to-noise ratio requirement					
Modulation type				BPSK	
Transmission type				TDM	
Transmission bit rate				1024.0 ksps	
Code rate				$^1/_2$ sequential	
Threshold BER				1×10^{-7}	
Ebt/No at threshold BER				3.0 dB	
BT product				1.14	
Conversion factor				-0.6 dB	
Equivalent threshold (C/N)				2.4 dB	
	Link performance summary				
	Clear weather	**Faded uplink**	**Faded downlink**	**Dual fades**	
Satellite					
Input backoff	18.9	18.9	18.9	18.9	dB
Output backoff	14.9	14.9	14.9	14.9	dB
% transponder power	12.7	12.7	12.7	12.7	%
% transponder bandwith	4.8	4.8	4.8	4.8	%
Receive site					
Noise temperature	353.1	353.1	460.1	460.1	K
Receive (G/T)	19.3	19.3	18.2	18.2	dB/K

continues

TABLE I (*Continued*)

	Link performance summary				
	Clear weather	**Faded uplink**	**Faded downlink**	**Dual fades**	
Performance					
Uplink C/N	19.4	19.4	19.4	19.4	dB
Downlink C/N	5.9	5.9	2.7	2.7	dB
Internal C/I	19.0	19.0	19.0	19.0	dB
External C/I	18.9	18.9	18.9	18.9	dB
Total C/N	5.3	5.3	2.4	2.4	dB
Threshold C/N	2.4	2.4	2.4	2.4	dB
Total margin	2.9	2.9	.0	.0	dB

	Uplink budget		
	Clear weather	**Dual fades**	
Transmit earth station			
Filter truncation loss	0.2	0.2	dB
HPA saturated output power	110.0	110.0	W
HPA output backoff	19.1	19.1	dB
Transmission line loss	1.5	1.5	dB
Antenna diameter	7.0	7.0	m
Operating frequency	14.250	14.250	GHz
Antenna peak gain	57.9	57.9	dBi
Antenna pointing loss	0.2	0.2	dB
Antenna polarization loss	0.1	0.1	dB
EIRP toward satellite	57.2	57.2	dBW
Transmission medium			
Elevation angle	25.0	25.0	degrees
Free space path loss	207.4	207.4	dB
Atmospheric absorption	0.3	0.3	dB
Uplink rain margin	0.0	0.0	dB
Propagation loss	207.7	207.7	dB
Satellite			
Receive signal power	−150.5	−150.5	dBW
Antenna gain per m^2	44.5	44.5	dBi/m^2
Receive power flux density	−105.9	−105.9	dBW/m^2
Saturation flux density	−87.0	−87.0	dBW/m^2
Input backoff	18.9	18.9	dB
Carrier-to-Noise Ratio			
EIRP toward satellite	57.2	57.2	dBW
Propagation loss	207.7	207.7	dB
Satellite receive G/T	2.0	2.0	dB/K
Boltzmann's constant	−228.6	−228.6	dBW/KHz
Demod noise bandwidth	60.7	60.7	dBHz
Uplink C/N	19.4	19.4	dB

Downlink budget			
Satellite			
Saturate EIRP	40.0	40.0	dBW
Output backoff	14.9	14.9	dB
Satellite EIRP	25.1	25.1	dBW

continues

TABLE I (*Continued*)

Downlink budget			
Transmission medium			
Elevation angle	25.0	25.0	degrees
Free-space path loss	205.8	205.8	dB
Atmospheric absorption	0.2	0.2	dB
Downlink rain margin	0.0	2.0	dB
Propagation loss	206.0	208.0	dB
Receive earth station			
Antenna diameter	1.8	1.8	m
Operating frequency	11.950	11.950	GHz
Antenna peak gain	44.8	44.8	dBi
Antenna noise temperature	40.0	40.0	K
Pre-LNA noise temperature	13.1	13.1	K
LNA noise temperature	300.0	300.0	K
Rain-induced noise	0.0	107.0	K
System noise temperature	353.1	460.1	K
Earth station G/T	19.3	18.2	dB/K
Carrier-to-noise ratio			
Satellite EIRP	25.1	25.1	dBW
Propagation loss	206.0	208.0	dB
Antenna point loss	0.1	0.1	dB
Antenna polarization loss	0.1	0.1	dB
Pre-LNA loss	0.2	0.2	dB
Earth station G/T	19.3	18.2	dB/K
Boltzmann's constant	−228.6	−228.6	dBW/KHz
Demod noise bandwidth	60.7	60.7	dBHz
Downlink C/N	5.9	2.7	dB

[a] Outroute refers to the signal path from the hub to the remote VSAT.

connection is based on the time-division-multiple-access (TDMA) technique. VSAT transmits only when it is required to do so. The network access is typically initiated through the slotted-ALOHA random access protocol, such that a VSAT within the network can contend for the network access privileges. All or selected VSAT stations are told by the hub that within a fixed designated time slot, VSAT can gain network access permission by transmitting a network access request. After the network access permission is granted to a VSAT, its network connection session bandwidth is assigned based on the demand-assigned-multiple-access (DAMA) method. In most networks, multiple frequency-division-multiple-access (FDMA) inroute channels are also available for network connection assignments. Figure 2 shows an actual VSAT network spectrum plot as seen by the Kuband geostationary satellite for a 512-kbps outroute signal from the hub station and a 128-kbps inroute signal from the VSAT station. This plot was taken from a standard VSAT video intermediate frequency (IF) port, and since the inroute transmission is intermittent, the spectrum plot was generated using the spectrum analyzer max-hold function.

As the geostationary satellite altitude is about 35,784 km above ground, the skewed signal path from each VSAT scattered on the ground has different time delays. Therefore, all VSATs must keep correct absolute timing as referenced at the hub station. Similarly, the transmit power of each VSAT as seen by the hub station may also be different because of the geographical satellite-coverage footprint variations and path losses. Thus, as far as inroute physical layer transmission is concerned, each VSAT must at least keep a configuration of its exact time delay and power settings with respect to the hub station. Naturally, there is a multitude of other station- and network-specific parameters that a VSAT must have in its station configuration, such as its network identification (ID), VSAT station ID, and the network-connection-related parameters.

VSAT network system design is considered as hub-centric, so that all references to system time, power, and network connections are viewed with respect to that of the hubstation. It is not surprising that a VSAT station is commonly referred to as a "remote." Thus, the network's single point of failure lies with the hub. A robust

TABLE II Inroute Link Budget Calculation Example for a Commercial 800-VSAT Station Network Belonging to a Hotel Chain

Operating parameters	
Channel	
Modulation type	BPSK
Transmission type	TDMA
Modem Transmission rate	256.0 ksps
Code rate	$1/2$ sequential
Threshold BER	1×10^{-7}
Threshold Ebt/No	3.0 dB
Transmit site	
Antenna diameter	1.8 m
Frequency	14.250 GHz
HPA size	1.0 W
HPA operating power	1.0 W
EIRP towards satellite	44.8 dBW
Receive site	
Antenna diameter	7.0 m
Frequency	11.950 GHz
LNA noise temperature	140.0 K
Satellite	
Receive G/T	−2.5 dB/K
Saturation flux density	-82.5 dBW/m^2
Saturated EIRP	43.7 dBW
Transponder bandwidth	43 MHz
Composite input backoff	10.0 dB
Composite output backoff	6.0 dB

Threshold carrier-to-noise ratio requirement	
Modulation type	BPSK
Transmission type	TDMA
Transmission bit rate	256.0 ksps
Code rate	$1/2$ sequential
Threshold BER	1×10^{-7}
Ebt/No at threshold BER	3.0 dB
BT product	1.14
Conversion factor	−0.6 dB
Equivalent threshold (C/N)	2.4 dB

	Link performance summary				
	Clear weather	**Faded uplink**	**Faded downlink**	**Dual fades**	
Satellite					
Input backoff	35.8	38.3	35.8	38.3	dB
Output backoff	31.8	34.3	31.8	34.3	dB
% transponder power	0.3	0.1	0.3	0.1	%
% transponder bandwith	1.2	1.2	1.2	1.2	%
Maximum number of carriers	83	83	83	83	
Receive site					
Noise temperature	193.1	193.1	337.7	337.7	K
Receive G/T	34.0	34.0	31.6	31.6	dB/K

continues

TABLE II (*Continued*)

	Link performance summary				
	Clear weather	**Faded uplink**	**Faded downlink**	**Dual fades**	
Performance					
Uplink C/N	8.6	6.1	8.6	6.1	dB
Downlink C/N	13.3	10.8	7.9	5.4	dB
Internal C/I	19.0	16.5	19.0	16.5	dB
External C/I	20.5	18.0	20.5	18.0	dB
Total C/N	6.9	4.4	4.9	2.4	dB
Threshold C/N	2.4	2.4	2.4	2.4	dB
Total margin	4.4	1.9	2.5	.0	dB

	Uplink budget		
	Clear weather	**Dual fades**	
Transmit earth station			
Filter truncation loss	0.2	0.2	dB
HPA saturated output power	1.0	1.0	W
HPA output backoff	0.1	0.1	dB
Transmission line loss	1.0	1.0	dB
Antenna diameter	1.8	1.0	m
Operating frequency	14.250	14.250	GHz
Antenna peak gain	46.4	46.4	dBi
Antenna pointing loss	0.2	0.2	dB
Antenna polarization loss	0.1	0.1	dB
EIRP toward satellite	44.8	44.8	dBW
Transmission medium			
Elevation angle	25.0	25.0	degrees
Free space path loss	207.4	207.4	dB
Atmospheric absorption	0.3	0.3	dB
Uplink rain margin	0.0	2.5	dB
Propagation loss	207.7	210.2	dB
Satellite			
Receive signal power	−162.9	−165.4	dBW
Antenna gain per m^2	44.5	44.5	dBi/m^2
Receive power flux density	−118.3	−120.8	dBW/m^2
Saturation flux density	−82.5	−82.5	dBW/m^2
Input backoff	35.8	38.3	dB
Carrier-to-noise ratio			
EIRP toward satellite	44.8	44.8	dBW
Propagation loss	207.7	210.2	dB
Satellite receive G/T	−2.5	−2.5	dB/K
Boltzmann's constant	−228.6	−228.6	dBW/KHz
Demod noise bandwidth	54.7	54.7	dBHz
Uplink C/N	8.6	6.1	dB

Downlink budget			
Satellite			
Saturate EIRP	43.7	43.7	dBW
Output backoff	31.8	34.3	dB
Satellite EIRP	11.9	9.4	dBW

continues

TABLE II (*Continued*)

Downlink budget			
Transmission medium			
Elevation angle	25.0	25.0	degrees
Free space path loss	205.8	205.8	dB
Atmospheric absorption	0.2	0.2	dB
Downlink rain margin	0.0	3.0	dB
Propagation loss	206.0	209.0	dB
Receive earth station			
Antenna diameter	7.0	7.0	m
Operating frequency	11.950	11.950	GHz
Antenna peak gain	56.9	56.9	dBi
Antenna noise temperature	40.0	40.0	K
Pre-LNA noise temperature	13.1	13.1	K
LNA noise temperature	140.0	140.0	K
Rain induced noise	0.0	144.7	K
System noise temperature	193.1	337.7	K
Earth station G/T	34.0	31.6	dB/K
Carrier to noise ratio			
Satellite EIRP	11.9	9.4	dBW
Propagation loss	206.0	209.0	dB
Antenna point loss	0.2	0.2	dB
Antenna polarization loss	0.1	0.1	dB
Pre LNA loss	0.2	0.2	dB
Earth station G/T	34.0	31.6	dB/K
Boltzmann's constant	−228.6	−228.6	dBW/KHz
Demod noise bandwidth	54.7	54.7	dBHz
Downlink C/N	13.3	5.4	dB

[a] Inroute is defined as the path from the remote VSAT to the hub.

and high-performance hub design is therefore imperative to the success of the VSAT network infrastructure construction.

III. VSAT HUB STATION DESIGN

A VSAT hub station is normally designed in a manner similar to other major standard "business service" satellite earth stations such as the Intelsat Standard C or E class earth stations. The Ku-band hub station antenna type used is typically a 9-m diameter parabolic, Cassegrain center-fed circular antenna. The antenna systems are usually electromechanically steerable with automatic satellite tracking control in antenna pointing. A deicing mechanism is almost always required in areas where snow and ice are a problem. The antenna uses linear orthogonal polarization. A 9-m diameter Ku-band hub antenna has gains of about 60.4 dB in the transmit band of 14.000 to 14.500 Ghz, and 59.2 dB in the receive band of 11.700 to 12.2 Ghz. The antenna pattern must meet the sidelobe standards set by FCC Rule 47 CFR Part 25.209.

The station-transmitter high-power amplifier (HPA) is on the order of 10 W, and the receiver LNA may sometimes have cryogenic cooling facilities. The hub transceiver is physically housed behind the antenna feed horn to minimize transmit power losses and receiver noise figure degradation. A block diagram that is representative of the design concept is depicted in Fig. 3. The IF of the transceiver is at 140 or 70 MHz, which is used by standard satellite modem equipment. Figure 4 shows a pictorial layout of a hub station infrastructure. An additional continuous-mode demodulator is sometimes present in the hub station. The continuous-mode demodulator serves as a means for the hub station to monitor the quality of its transmit signal. Most VSAT hub stations use standard satellite modem equipment, which is connected to network control computer clusters. Although standard satellite communication systems prefer the use of QPSK modulation for its optimum power and bandwidth performance, VSAT systems mainly employ the BPSK modulation technique. It is easy to realize from the VSAT link budget that a VSAT network has fundamental power limitations imposed by the space segment. Thus, spectrum efficiency is not as

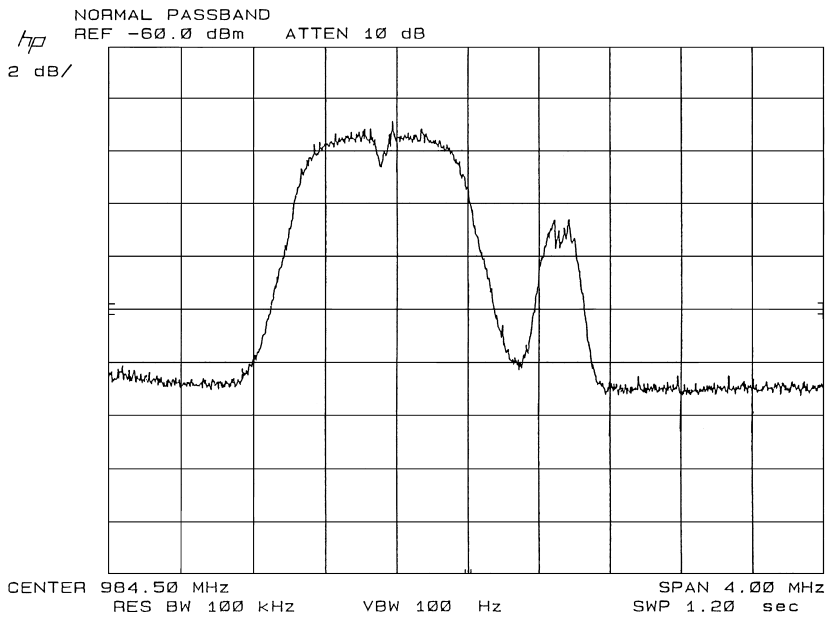


FIGURE 2 This is a spectrum analyzer plot taken at the video IF port of a VSAT showing the outroute spectrum of an 1.024 Msps BPSK signal supprting a 512 kbps data rate at rate $^1/_2$ convolution coding. A single inroute BPSK signal of 256 ksps is also shown here that provides 128 kbps data rate at rate $^1/_2$ convolution coding. The inroute spectrum plot is captured by using the max-hold function of the spectrum analyzer because the inroute signal has a bursty transmission nature.

important as the robustness and simplicity of the modulation technique. However, in special cases such as SCPC or mesh network configurations, the link budget may be set up in such a way that this argument does not apply.

Although a VSAT hub station is often installed and constructed with its own dedicated building structure, its cost impact on the overall network deployment cost is minimum. When the hub station cost is amortized over the

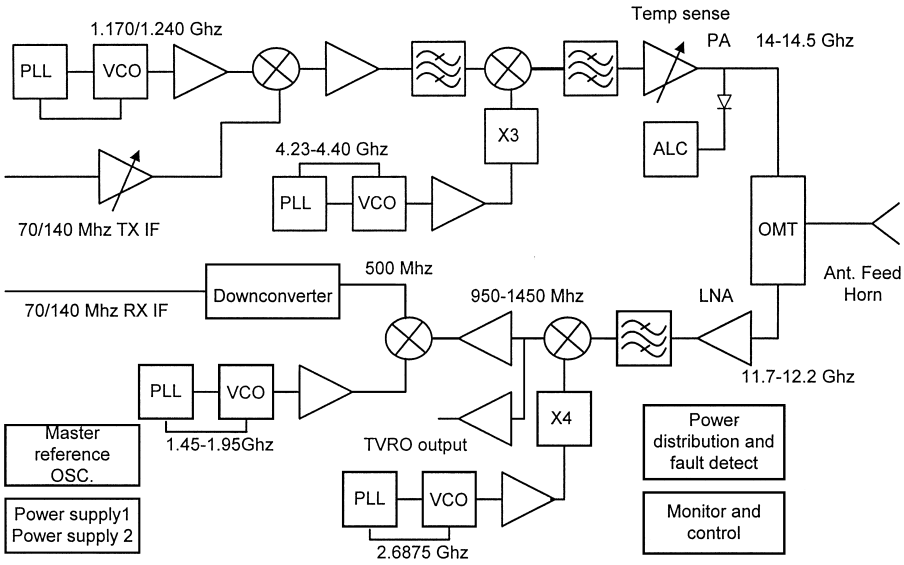


FIGURE 3 This is a block diagram of VSAT hub transceiver design. The IF interface of the transceiver is typically at 70 Mhz or 140 Mhz, similar to standard "business-service" class of satellite earth stations. Full fault moniter and control facilities are normally available.

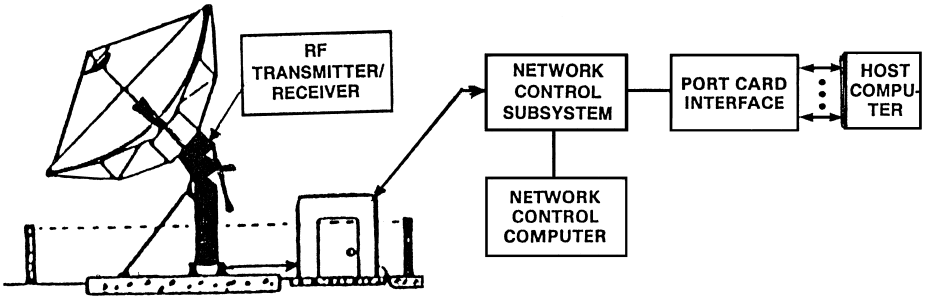


FIGURE 4 This is a pictorial view of the layout of typical VSAT hub equipment. A building is often present to house the modem and baseband equipment. Network control clusters are normally resided in the same building.

size of the network, its cost contribution is in general not the most significant cost item when compared with the cost of the individual VSAT station.

IV. VSAT STATION DESIGN

In a VSAT network, the biggest cost contribution to the network deployment financial model is the price of the VSAT station. Therefore, the design used in VSAT stations is very cost sensitive. Ever since the genesis of VSAT technology in the early 1980s, the transceiver design used in VSAT stations has been highly integrated. The design called for a combined transceiver and antenna feed horn structure, which is placed at the VSAT antenna feed point.

The majority of the Ku-band VSAT antennas used are of the 1.8-m diameter, parabolic, offset-fed prime focus variety. There is a good reason for this choice. In 1984, the FCC adopted requirements to protect 2° geostationary satellite parking spacing. In these requirements, VSAT antenna radiation pattern must comply with a strict antenna pattern sidelobe profile in accordance with the 29-25 log θ dBi bore-sight rule for $1.6^\circ \leq \theta \leq 7^\circ$. More specifically, for $7^\circ \leq \theta \leq 9.2^\circ$ the sidelobe must not exceed 8 dBi and again 32-25 log θ dBi for $9.2^\circ \leq \theta \leq 48^\circ$. This ruling limits the choice of antenna type. Indeed, a 1.2-m parabolic, offset-fed prime focus antenna will be in violation of this rule between $\theta = 1.00^\circ$ and 1.40°. The network operators must therefore convince the authorities to allow the legal deployment of such antennas by using lightly spread-spectrum outroute signals and inroute signals with reduced data speed. One may have observed the common use of rectangular antennas in VSAT networks. A rectangular antenna has a knife-shaped beam, with the thickness of the knife shape corresponding to the larger dimension of the antenna and the width of the knife shape to the shorter dimension of the antenna. Since one is only concerned with the adjacent satellite interference in a geostationary satellite orbit scenario, one can cause the thickness of the knife-shaped beam to be perpendicular to the satellite orbit plane to satisfy this requirement. A rectangular antenna gives the visual impression of being smaller than a parabolic antenna of the same dimensions. It is also customary to list only the small dimension of the antenna in the literature.

Antenna deicing mechanisms are very commonly installed on antennas in places where there is snow and ice in winter. In the early days of VSAT technology, there was also a concern for the undetermined link-budget burdens imposed by wet antenna-reflector surfaces during rainfall. A hydrophobic chemical was used to keep the antenna surfaces dry. Today, there is a better understanding of this phenomenon (Cheah; 1993).

There are two approaches to VSAT outdoor-unit (ODU) design. One is to output the bidirectional transmit and receive signals at L-band IF to the indoor unit (IDU). In this way, the delicate modem and baseband electronics can be safely kept away from the harsh natural elements outside. It is also a lower-cost design because it minimizes the number of industrial-grade circuit components used. However, it suffers from one major drawback. There is a finite limitation on the length of the cable that can be used in between ODU and IDU, and therefore it presents an installation problem. The length of the cable is determined by the dynamic range of the respective IF automatic level control (ALC) circuit. At L-band IF, the practical length of the cable allowed is about 100 m. This is clear from the typical Ku-band ODU and IDU block diagrams shown in Fig. 5.

Another design approach is to output baseband signals from the ODU. In this way, the cable from the ODU to the IDU can be of any length, with only the addition of baseband signal repeaters in between where necessary. However, although this design was common in the early days of VSAT technology, it is not popular because of its cost and the high instances of ODU failure due to the complexity of the designs. A representative block diagram is shown in Fig. 6.

One interesting element of the VSAT station design that needs to be highlighted is the recovery of the frequency reference of the hub. In TDMA systems the transmission

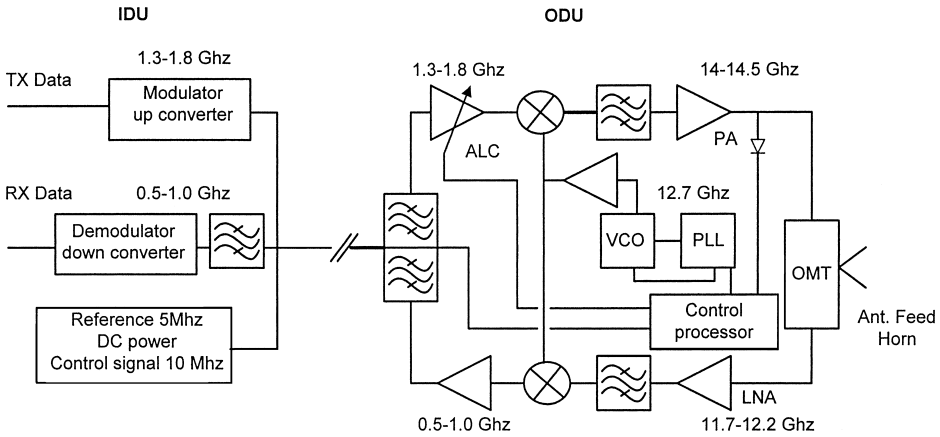


FIGURE 5 A modern VSAT transceiver block diagram. The ODU and the IDU are linked by a coaxial cable at L-band frequencies. In this configuration, the IDU has an L-band RF module, and the L-band cable has an average length of about 100 m.

burst signal from the VSAT station must have a very good transmit frequency accuracy agreement with that of the hub's receiver. Typically, the accuracy is required to be in the order of plus or minus a few hundred hertz. Thus, the hub burst demodulator does not need very long signal preambles from the VSAT to acquire the transmission. To place such an accurate frequency reference in every VSAT throughout the network would be prohibitive in cost. Most VSAT designs solve this problem by recovering the frequency reference through the hub's TDM data clock stream. In effect, it forms a frequency lock loop spanning the satellite link. This method (Cheah, 1994) also bypasses all the frequency inaccuracy that exists within the satellite and the VSAT receiver chain. Since the VSAT station derives its frequency reference from the hub, the master frequency oscillator of the hub also is not required to be unduly accurate. The required accuracy of the hub's master frequency oscillator is simply to meet the space segment vendors' requirements.

V. VSAT SPACE SEGMENT

Today, the space segment vendors are very efficient and well set up to cater for VSAT needs. There are many satellites choices with sufficient space segment capacity. A list of the available satellites can be found in Cheah (1999). The space segment user pays for the spacecraft resources according to the extent of usage. Each satellite transponder has fixed power and bandwidth resources. Thus, the user

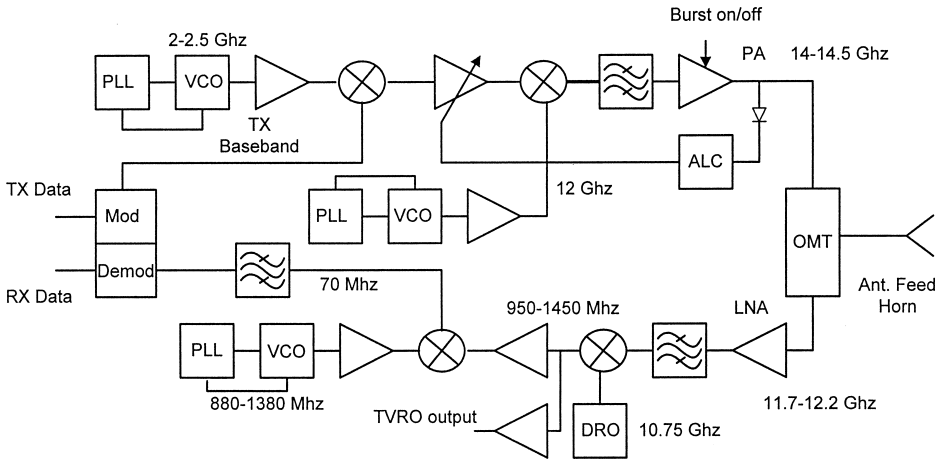


FIGURE 6 A VSAT transceiver block diagram. The video IF output uses a dielectric resonator oscillator (DRO) as a local oscillator (LO). The ODU and the IDU are linked at baseband signal level. The IDU does not contain any RF components. In this configuration, it is possible to repeat the baseband signals easily, and therefore the inter ODU/IDU cable length can be unlimited.

pays for the percentage of the transponder power required if the network is power limited, or for the percentage of bandwidth used if it is bandwidth limited. It is the user's responsibility to prepare an accurate operation link budget so that the appropriate space segment resources can be purchased. In a typical VSAT network, it is the transponder power usage for which the user will be charged. Once the target satellite is known, then an accurate link budget such as those of Tables I and II can be finalized.

A Ku-band satellite operates mainly in linear polarization with some exceptions, such as the USDBS series of satellites that adopted circular polarization. A similar set of space segment resource is also available in the cross-polarization direction. A VSAT station is particularly prone to interfere and be interfered with when the polarization orientation is poorly aligned during installation. The polarization discrimination is determined solely based on the performance of the orthogonal mode transducer (OMT) at the antenna feed horn, and its physical orientation with respect to the satellite's polarization in space. However, the polarization orientation angle can be accurately calculated.

Localized suborbital figure-of-8 satellite movement is not taken into consideration by the VSAT station because of the obtuse nature of the antenna radiation pattern at bore-sight. This is a consequence of the small physical size of the VSAT antenna.

For VSAT applications using Ku-band satellites, there may be a need to avoid transponder frequencies lower than 14.050 Ghz and frequencies that are higher than 14.400 Ghz. This is to avoid the necessity for site-by-site frequency coordination procedures because these frequencies are shared by the terrestrial microwave facilities in some areas.

VI. VSAT APPLICATIONS

Early VSAT systems were used to disseminate financial data from main financial institutions to branch offices. Often, the network was used in the receive-only mode. Later, the most notable large-scale two-way VSAT networks were deployed in the gas exploration and supermarket retail industries. Indeed, at that time, there was a memorable advertisement slogan by one of the pioneering VSAT network users in the supermarket retail industry declaring that "We sell a bar of soap through satellite communications."

Today, almost all the credit-card transactions in gas stations are transported through VSAT networks. Car dealerships rely on VSAT networks to provide up-to-the-minute car inventory availability throughout the country very effectively. Banks and insurance companies, drug stores, hotel chains, car rental businesses, oil and gas pipelines, and others similarly integrate VSAT data transport systems into their computer network successfully.

Voice connectivity is also supported by VSAT networks at the expense of data traffic in some cases. Isochronous voice data traffic is not really a very efficient use of the packet data transport backbone provided by the VSAT network. In addition, broadcast video can also be easily provided through the low-noise-block (LNB) output of a VSAT transceiver if required.

VII. WEAKNESSES OF VSAT NETWORKS

VSAT network connection is pivoted at the space segment. Unfortunately, in the past, there were a few instances where rogue and therefore harmful satellite interference existed in the space segment for some short periods of time. Consequently, major disruption to the operation of the network affected occurred. Although each satellite interference event is carefully and seriously studied and investigated by all concerned (Cheah and Davis, 1997), it remains a potential problem that VSAT network operators must plan for. There is technology available to estimate the location of the spurious interference source. For example, the signal time difference of arrival (TDOA) from the adjacent satellites can be used (Chestnut, 1982). It is also possible to use the differential Doppler frequencies or frequency difference of arrival (FDOA) methods or a combination of all these methods to pinpoint the location of the offender. It is common knowledge that space-segment vendors are actively employing such technology to safeguard their assets. Nevertheless, a good VSAT network must have suitable, temporary backup communications resources to overcome outages such as this. Temporary overflow backup terrestrial lines can be put in place to service outages or occasional data traffic congestions. Dial backup modems that use the standard public switch telephone network (PSTN) may be another option. Most VSAT station designs include some types of catastrophic failure recovery state machines, whereby a VSAT will have a predetermined recovery algorithm to search the satellite for a switched outroute frequency if it loses connection through the original outroute channel. In this way, members of a VSAT network will automatically rejoin the network even if the hub station makes an unannounced outroute frequency change because of an unacceptable interference episode or a catastrophic transponder failure.

Apart from the space segment, the VSAT hub station is also a single point of failure. The obvious potential problem is electronic component failure. The security of remedy can be suitably parlayed through redundancy strategies. Natural disasters such as earthquakes, floods,

hurricanes, or tornadoes are major concerns. Such natural disasters may present a prolonged network outage possibility. A technique of seamlessly providing a geographically remote backup hub function is available. The backup hub station monitors the main hub's transmission in the background. It copies all the traffic of the main hub and the VSAT without turning on its own transmitter. When the main hub's station's signal fades for more than 3 dB because of some ground-based catastrophic failure, the backup hub will commence transmission at the exact same frequency as the main hub. All the active data connection sessions among the VSATs can be maintained without interruption. Such a geographical remote hub redundancy product is available in the VSAT market.

Interference at VSAT stations also causes localized network outages. The interference can be electromagnetic or physical in nature. For example, there have been recorded instances of Ku-band interference generated by motorists using faulty radar detectors. In addition, physical interference such as willful vandalism is quite common. Physical interference inevitably required maintenance service calls that translate to a lengthy network down time.

Periodic maintenance of the VSAT is imperative. Qualified personal must check the accuracy of the antenna pointing. The cleanliness of the antenna feed horn radome must be maintained. Other station physical integrity should be routinely inspected. Only in this way can local network robustness be ensured.

VIII. OTHER SPECIALIZED VSAT NETWORKS

Almost 80% of the VSAT network population is of the star-connected variety. There are some SCPC VSAT networks where two VSAT stations communicate with each other through a fixed channel dedicated to this communication. In this case, the VSAT must be suitably equipped so that it can satisfy all the regulatory requirements. For a network of multiple VSAT with this capability, it is convenient to network all the stations in a mesh network. To do so, each VSAT must also have the network control functionality so it can set up and tear down a data connection session. The best example for a full function mesh networking using VSAT technology is the telephony earth station network. The telephony VSAT mesh network is designed to service voice connections in mesh connections. In general, they have much bigger antenna structures so as overcome the mesh connectivity requirements. Such an earth station is usually much higher in cost than a standard VSAT station.

SEE ALSO THE FOLLOWING ARTICLES

COMMUNICATION SYSTEMS, CIVILIAN • NETWORKS FOR DATA COMMUNICATION • SATELLITE COMMUNICATIONS • WIRELESS COMMUNICATIONS

BIBLIOGRAPHY

Chakraborty, D. (1988). "VSAT communications networks—an overview," *IEEE Commun. Mag.* **26,** 10–24.

Cheah, J. Y. C. (1993). "Wet antenna effect on VSAT rain margin," *IEEE Trans. Commun.* **41,** 1238–1244.

Cheah, J. Y. C. (1994). "Frequency reference in VSAT," *IEEE Trans. Commun.* **42,** 233–236.

Cheah, J. Y. C. (1999). "Practical Wireless Data Modem Design," Artech House Publishers, Boston.

Cheah, J. Y. C., and Davis, M. E. (1997). "Tone interference in VSAT spectrum," *IEEE Trans. Commun.* **45,** 1035–1038.

Chestnut, P. C. (1982). "Emitter location accuracy using TDOA and differential Doppler," *IEEE Trans. Aerospace Electron. Syst.* **AES-18,** 214–218.

FCC51 (1986). "Declaratory order in the matter of routine licensing of large networks of small antenna earth station operating in the 12/14-Ghz frequency bands," *Fed. Reg.* 15067.

Lyon, D. L. (1985). "Personal computer communications via Ku-band small earth stations," *IEEE J. Select. Areas. Commun.* **SAC-3,** 440–448.



Wireless Communications

Michele Zorzi
Università di Ferrara

A. Chockalingam
Indian Institute of Science

I. The Key Technical Challenges
II. Current Wireless Cellular Systems
III. Wireless Data Systems
IV. Third Generation (3G) Cellular
V. Research and New Technologies for 3G and Beyond
VI. Further Reading

GLOSSARY

Ad hoc wireless network A collection of wireless mobile hosts forming a temporary network without the aid of any established infrastructure of centralized control.

Cellular system A mobile radio system in which the assigned frequency spectrum is divided into discrete channels which are assigned in groups for use in different geographical areas (called cells). Several cells are grouped to form a service area.

CDMA Code division multiple access, a modulation and multiple access scheme based on spread-spectrum techniques. it allows universal frequency reuse.

Frequency reuse Allowing the same radio frequency channel to be used by different cells in the same service area. Reuse requires sufficient cell separation to provide adequate signal-to-interference ratio. Reuse helps to increase the cellular system capacity (number of mobile subscribers supported).

Multiple access Method by which different mobile transmitters share a common radio frequency channel to transmit information. Popular multiple access techniques include frequency division multiple access (FDMA), time division multiple access (TDMA), and code division multiple access (CDMA).

Third generation (3G) wireless Digital cellular systems having greatly enhanced performance and service characteristics including multimedia services (voice, data, high quality image, video, etc.) at high data rates (up to 2 Mbps) with different quality of service requirements.

Wireless data Computer-to-computer communications through wireless medium.

EVEN THOUGH the radio was invented more than 100 years ago, it has not been until 10–15 years ago that widespread use of wireless communications had begun. While before then it was mainly used by police and military, now the capability of communicating without being constrained to a fixed physical location is within everyone's reach. The booming industry which has been growing at such an astounding rate in the past couple of decades

is a very good indicator of how much such a capability meets the human desire for freedom.

Besides being a very exciting and profitable business, wireless and mobile communications has been an extremely rich field for research, due to the many difficulties that the wireless environment presents, and due to the ever increasing users' demand for more, newer, and better services. People want to benefit from wireless services, and as these become more and more widespread, increasing expectations challenge manufacturers and providers to continuously search for new solutions.

From the technical point of view, the past few years have seen tremendous advances in the research, development, and design of mobile radio systems, and many more are expected in the near future. The purpose of this article is to give a brief overview of this exciting field and to provide some basic information about the problems inherent to the wireless environment, the solutions on which the implementation of the systems in use today are based, and the research ideas which are being developed and will be the foundation of the telecommunications systems of tomorrow.

I. THE KEY TECHNICAL CHALLENGES

The idea of being able to communicate with other people or with machines (e.g., databases or video servers) without being constrained to a fixed location is of course very appealing. Unfortunately, mobility introduces new management issues and the mobile wireless environment is very hostile when compared to classic telecommunications systems. In particular, radio channels are inherently prone to severe impairments (especially when the transmitter and/or the receiver are moving) and have limited bandwidth. In this section, we describe the key features which make the wireless environment so challenging.

A. Channel Unreliability

Probably the most notorious characteristic of wireless channels in mobile communications is their unreliability. Due to a number of physical factors, signals propagating through such channels are subject to severe impairments. The effect of propagation on the transmitted signal can be adequately described by three phenomena, namely, path loss, shadowing, and multipath fading.

Path loss accounts for the signal attenuation due to the physical distance, r, between transmitter and receiver. In free space, the received power is proportional to r^{-2}, whereas in groundwave propagation there is an additional effect due to the combination of the direct (free space) propagated wave and its replica as reflected by the earth's surface. In this case, the path loss behavior is better described by the function $r^{-\eta}$, where η depends on a number of properties of the environment, and usually takes values between 2 and 5. More accurate models account for other effects, by introducing some correction factors in the relationship between received power and distance to model, for example, effects similar to guided propagation (as happens along streets with tall buildings). Such models are usually employed in system simulations, and their parameters can be chosen by direct measurements of the environment to be studied.

Another macroscopic effect, i.e., an effect which does not change much over distances on the order of many wavelengths, is the shadowing effect (sometimes called shadow fading) which, as the name suggests, accounts for obstructions along the propagation path (e.g., building or trees), which may cause significant signal attenuations. As these attenuations combine with each other, shadowing is often modeled as having Gaussian distribution when path loss is expressed in dB. This model (also called *log-normal shadowing*), has been found to match field measurements very well.

On a much finer scale, signals are subject to multipath fading. Due to various objects and reflectors, the signal does not directly propagate to the receiver, but rather it is scattered in many directions so that multiple versions of it, coming from different angles and with different attenuations and delays, are received. If the delay differences among the various reflections are much smaller than the time scale at which the signal is detected (e.g., the bit duration), then the channel impulse response is very short (in time), and the effect of multipath in this case is only an additional attenuation (often modeled as a Rayleigh or Rician random variable). Such a situation corresponds to the transfer function of the channel being essentially constant over the signal bandwidth, and is called *flat fading*. If, on the other hand, the delay distribution on the different paths is spread out (e.g., it exceeds the bit duration), then the channel changes the spectral shape of the transmitted signal, which is therefore distorted. As different frequencies may undergo different attenuations, this case is called *frequency-selective fading*, and is typical of wideband signaling. For digital communications, some sort of equalization is usually needed in this case. Also, the fact that not all spectral components fade together can be exploited as a form of frequency diversity.

Since in a wireless system terminals are allowed to move, the channel impulse response changes in time in general. Even in the case in which terminals do not move, the environment around them may change (e.g., vehicles or people passing by). Path loss and shadowing are relatively constant over significant distances (tens of meters or more), whereas multipath fading can usually be

considered independent at distances on the order of the wavelength (which is equal to the speed of light divided by the carrier frequency, and is $\lambda = 33$ cm at 900 MHz), and therefore changes very rapidly. This means that the channel, besides being subject to significant impairments, is also very unstable and sometimes may even suddenly disappear. This makes wireless communications very different from communications over more stable and predictable channels such as cables and fibers or even microwave links. Stabilizing the radio channel and overcoming problems posed by the propagation environment is a major task in the design of mobile communications systems.

Typical techniques which are used in today's systems to counteract the channel behavior are power control and diversity. Power control is based on the concept of trying to boost the power if the channel is currently undergoing large attenuations and to reduce the power when the channel experiences good conditions. This ideally results in all users transmitting just as much power as they need in order to meet their quality objectives, and therefore minimizes the interference level caused to others, under quality of service (QoS) constraints. Implementation of the power control mechanism requires knowledge of the channel attenuation. In open-loop power control schemes,* the transmitter autonomously estimates the channel attenuation (either measuring the received signal in full duplex communications, or through the use of a pilot channel) and adjusts the transmitted power in order to compensate for such attenuation. In closed-loop power control, based on the quality of the received signal, the receiver sends feedback to the transmitter indicating whether the power should be increased or decreased. Open-loop power control is easier to implement but cannot usually compensate for multipath fading, which is not reciprocal if the transmitted and received signals occupy different frequency bands.** Therefore, in order to be able to compensate for multipath, closed-loop power control is usually envisioned. On the other hand, since multipath fading is rapidly varying, this results in potentially inaccurate channel estimation and significant overhead due to the need for frequent updates.

A classic means of counteracting multipath fading is through diversity, which is based on the fact that multiple copies of a signal received at antennas sufficiently apart from each other (a wavelength is usually considered sufficient, at least at the mobile terminal) experience almost independent fading, so that it is very unlikely that all antennas receive weak signals at the same time. More advanced diversity systems try to optimally combine all received signals rather than just selecting one of the signals.

Robust modulation techniques can also be used to maintain transmission quality even under severe propagation conditions. A well-known example for digital communications is spread-spectrum modulation, where signal-processing techniques are used at the transmitter to spread the signal power over a wide band and at the receiver to despread it. At the same time, the power spectral density of other spread-spectrum users is essentially unchanged, whereas the power of narrowband interferers is spread over the whole bandwidth. Therefore, when filtering over the intended signal bandwidth after despreading, the power interference is greatly reduced. The robustness of this technique to interference comes at the expense of bandwidth expansion.

Finally, another well-established means to fix the channel unreliability in digital communications is through the use of error control techniques, whereby redundant symbols are added to the information stream to be transmitted, and are used at the receiver for enhanced detection. Error correction schemes, traditionally designed for channels with additive white Gaussian noise, are applied in this environment jointly with interleaving, whereby the bits are scrambled prior to transmission and descrambled at the receiver, thus breaking the error correlation inherently present because of the channel memory. The interleaving/deinterleaving operation introduces some delay which effectively limits the extent to which correlation can be removed: in the presence of fast varying channels interleaving usually works very well, whereas for slow channels it is basically ineffective (but in this case very accurate power control can be performed due to high channel predictability).

B. Limited Bandwidth

Another important difference between wireline and wireless networks is the amount of bandwidth available. One reason for this is that practical communications schemes today operate at UHF frequencies (typical bands are 900 MHz and 1.8 GHz), and additional frequency bands being considered for future systems are mostly in the 1–6 GHz range (even though 17 GHz has been designated for wireless LANs in Europe, and there are proposals for systems at frequencies up to 60 GHz). The amount of bandwidth available in these ranges is not large, since the radio spectrum itself is an inherently public resource, and is already very crowded due to the presence of other services (e.g., broadcast TV, military communications, and point-to-point radio links). Therefore, wireless systems are always limited by interference, which dictates the

*The term open-loop is used here to indicate that there is no *explicit* signaling in the return direction to control the transmitted power, even though a return channel is in fact present and used to estimate the attenuation.

**This approach to separating transmitted and received signals is referred to as frequency division duplexing (FDD).

amount of bandwidth available and imposes restrictions on how it can be used, whereas interference effects in wireline networks, although present (e.g., crosstalk), are less significant in limiting the capacity. Important consequences of this fact are (1) the limited bandwidth available for wireless services, which constrains both the capacity (i.e., the number of users the system can accommodate) and the user's capabilities (i.e., the bandwidth which may be instantaneously available to a user) and (2) the need to conform to regulatory constraints. The fact that the transmission medium in radio communications is inherently public also raises concerns about its security and privacy. Means need to be provided to guarantee that telephone conversations or data transactions cannot be eavesdropped (matching at least the security achieved in wireline networks). In addition, in this environment, privacy about the user's location and equipment security (e.g., to avoid unauthorized use of cellular telephones or of billing identification numbers) needs to be ensured.

Because of the shortage of the radio spectrum, a major research objective toward improved design of wireless networks has been to provide more bandwidth-efficient techniques, in order to increase the number of users who can be accommodated for a given system bandwidth. A major breakthrough in this sense was the invention of the cellular concept, devised at Bell Laboratories in the 1970s. Before then, the number of available channels was very limited, because of the limited spectrum available. Cellular systems are based on the same concept used for many years in broadcasting, where the same channel can be used in noninterfering geographic locations. This allows a capacity increase which is in principle unlimited, even though other issues make it infeasible to reduce the cell size beyond a certain limit.

Typical cellular systems today are based on circuit switching, which has been a very successful paradigm for speech communications. The total system bandwidth is split into a number of channels, and these channels are in turn assigned to cells, according to the frequency reuse plan, if any. Therefore, each cell has a number of channels, and cells using the same channels are separated geographically by a sufficiently large distance, thereby causing small interference to each other. Cells using the same radio channels belong to the same co-channel set, and the number of distinct co-channel sets is sometimes called *cluster size* (see Fig. 1). How the system bandwidth is subdivided into channels depends on the multiple access technology of choice. The early systems used frequency division multiple access (FDMA), in which the system bandwidth is sliced in the frequency domain to achieve orthogonal (nonoverlapping in frequency) channels. With the introduction of digital transmission in wireless communications systems, other techniques became possible,

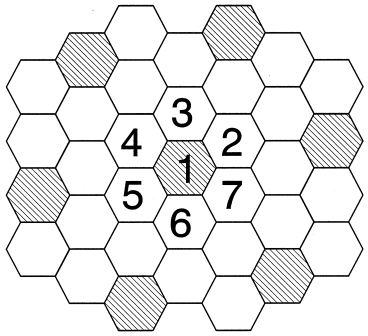


FIGURE 1 Example of cellular layout with seven-cell frequency reuse plan. Cells with different numbers use different radio channel sets. The shaded cells all use channel set number 1.

most notably time division multiple access (TDMA) and code division multiple access (CDMA). In the former, the time axis has a framed structure, and different channels use the same bandwidth but at different times, thereby achieving orthogonality in the time domain. In CDMA, spread-spectrum modulation is used to allow multiple users to communicate simultaneously in the same radio channel, and the separation is obtained through signal-processing techniques: if all signals are synchronously received, and in the absence of multipath, it is possible to achieve perfect orthogonality in the code domain, whereas some residual interference will be present if multipath is present or synchronization is not possible (due, for example, to different propagation delays). Hybrid multiple access schemes can also be envisioned, and are actually used in many practical system implementations. A very important advantage of CDMA with respect to TDMA and FDMA is its inherent robustness to interference, which makes it possible to use the same frequency band in all cells. This eliminates the need for frequency planning (which is a major undertaking, especially in systems with small cells), and achieves higher capacity, since the fact that the same frequency can only be used in distant cells results in a significant capacity penalty in TDMA and FDMA, where the bandwidth assigned to each cell is only a fraction of the total.

A major improvement in the spectrum utilization efficiency has been achieved with the introduction of digital communications systems. Digitized speech can be easily compressed, so that, for example, the 30 kHz necessary for a single FM analog voice channel (used in the early systems) can support three digital voice channels with data rate of 16 kbps. The original pulse code modulation technique, where voice signals were sampled and quantized, has been successively refined, based on exploiting the large amount of redundancy of the source, and schemes requiring fewer and fewer bits per second to reliably transmit a single telephone call have been developed. An even more dramatic compression rate can be achieved

by vocoders which, instead of compressing the sampled version of the actual speech waveform, model the mechanics of how the sounds are produced as linear filters, and transmit their parameters which are then used at the receiver to generate the same sounds. The currently used VSELP and RPE belong to this latter class.

The increased compression rates, along with the use of bandwidth-efficient digital modulation schemes such as QPSK or Gaussian MSK, make it possible, for the replacement of each 30-kHz analog channel with three TDMA channels, thus tripling the system capacity. An even larger gain may be obtained when using CDMA, because of the possibility of reusing the spectrum everywhere, so that the capacity gain offered by CDMA with respect to the early systems is estimated to be on the order of ten to twelve in circuit-switched systems.

C. Support for Mobility

The third important technical challenge besides channel unreliability and limited spectrum is the need to support user mobility. Even though mobility in wired networks (as in Mobile IP) and wireless systems with no mobility (as in some Wireless LAN and Wireless Local Loop applications) have been considered, the issue of mobility remains largely an essential feature of wireless mobile networks. A whole new approach needed to be developed in order to locate and track users and to provide a means to change the point of attachment to the network, even during an active connection. Implementation of these capabilities requires a considerable amount of signaling and network management software.

Two basic problems need to be solved to guarantee effective mobility management, i.e., location and handoff. The location problem arises because the system needs to know where a user is at any time, in order to be able to correctly forward incoming calls. In current cellular systems, the network is divided into service areas, each composed of a number of cells managed by a central unit, called Mobile Switching Center (MSC).* Each MSC has two databases, called Home Location Register (HLR) and Visitor Location Register (VLR). The HLR contains information regarding users who are registered as subscribers in the area, whereas the VLR contains information regarding users who are registered as subscribers somewhere else but happen to be roaming in the area. Every time a user registers as a roamer in a new area, some signaling takes place between the MSCs involved, and the databases are appropriately updated. Basically, the information contained in those registers allows the network to locate the user, i.e., to identify the set of cells where the user can be found at any given time. In fact, every time the network needs to locate a user, the HLR at the user's home MSC will be looked up. If the user is not in that area, the HLR contains information on where the user can be found. If an incoming call for that user arrives, it may be redirected to the MSC of the area in which the user can be found. The MSC will then instruct the central transmitters, called base stations (BS), serving all cells in that area to send a paging message for the user, who will reply to the BS of the appropriate cell. This response completes the location procedure, and the call setup can be completed at the designated BS.

In order to know where a terminal is at any time, the network must be able to track it while it is moving. For this purpose, signaling messages need to be frequently sent to update the network's information about the user's location. In particular, the system is divided into location areas, and every time the user crosses the boundary of one of those, a location update message needs to be sent. Large location areas result in little update signaling but large paging overhead (all cells in the area must send a page for the user), whereas the opposite is true for small location areas.

The above location procedures take place when the user is not sending information (even though the terminal is, but just for signaling purposes). However, users may move while communicating as well. The second major problem to be solved in a mobile communications system is therefore to handle ongoing calls when, because of mobility and/or changing channel conditions, the user needs to change its point of attachment to the network. A typical situation, although by no means the only possible, occurs when a user moves away from the BS at which the call was established, going toward another BS. At some point, the attenuation experienced on the path toward the first BS may become too severe, to the extent that the transmission quality is no longer adequate. A system with no support for mobility would drop the call at that point. In cellular systems, however, this situation would trigger a handoff procedure, whereby the mobile terminal and/or the serving BS look for an alternative BS which can guarantee adequate transmission quality. If no such alternative can be found, the call is dropped; otherwise a handoff occurs, i.e., the connection to the old BS is released and a connection with the new BS is established.

The details of the handoff procedure depend on the system implementation. In general, however, it involves periodic measurements of the quality (typically, the signal strength or the signal-to-interference ratio) of the channel currently being used and of the paths to a number of candidate BS. Whenever the need arises, those measurements make it possible to determine which BS should be chosen as the destination of the handoff procedure. The decision about when a handoff is performed and which

*Different names are used in different systems, but the basic functionalities remain essentially the same.

BS is selected can be made either by the network alone, or with the cooperation of the mobile terminal (popularly known as the Mobile Assisted HandOff, MAHO).

II. CURRENT WIRELESS CELLULAR SYSTEMS

The following are the main elements of a typical cellular system (see Fig. 2):

- Mobile terminal, which contains a radio transceiver and a processor
- BS, which manages communications to and from the users belonging to its own cells, and is composed of a radio part and a base station controller;
- MSC, which controls a number of BS (typically a few tens), provides additional functionalities such as call management and diagnostics and fault recovery, and is the interface to and from the public switched telephone network (PSTN)
- Databases, such as the already mentioned HLR and VLR, plus the equipment identity register and authentication center, used for security and authentication purposes.

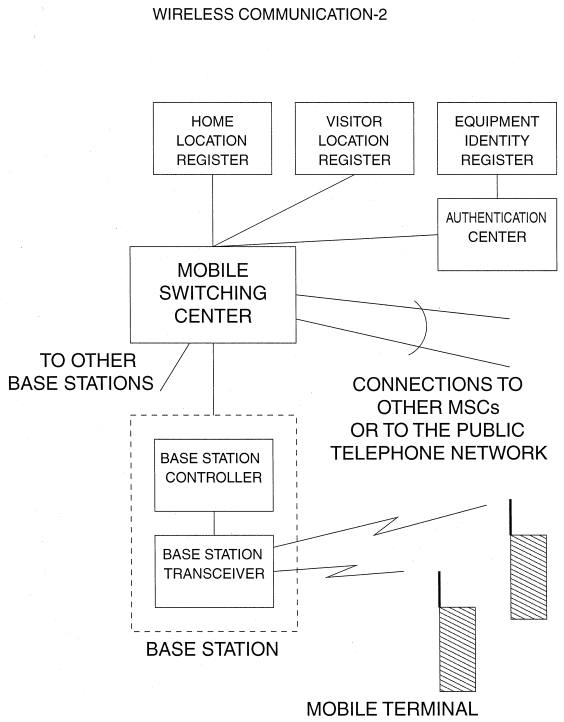


FIGURE 2 Main components of a cellular system and their relationships.

Also, the following control channels are present in all systems:

- Broadcast channels (base-to-mobile), where general control information is transmitted to all mobiles
- Paging channels (base-to-mobile), used to notify a mobile user of an incoming call
- Random access channels (mobile-to-base), used for mobile-initiated call requests

These elements provide essential communications services and network management functionalities, which enable proper system operation. In particular, they provide location and mobility management. Therefore, although in different forms and possibly with different names, they are present in all cellular systems.

A. First Generation Cellular Systems

The early cellular systems were based on analog FM. Examples of such systems are the Advanced Mobile Phone Service (AMPS) in the United States and the Total Access Communications System (TACS) in Europe. The AMPS system operates in the vicinity of 850 MHz. Two competing operators are allocated 25 MHz of bandwidth each, corresponding to 833 radio channels, each occupying 30 KHz. Because in these systems the modulation format for the voice channels is analog, FDMA must be used for multiple access. With a seven-cell frequency reuse plan (which guarantees adequately low interference), this results in about 60 full-duplex voice channels per cell.

First generation systems provide supervisory signaling during voice transmission, to verify proper connection between BS and mobile terminal. This is achieved by sending tones at frequencies above the audio bands. In addition, signaling capability is provided (e.g., for use during handoffs) by means of the *blank-and-burst* technique, in which bursts of data are sent for short (almost unnoticeable) durations while speech is muted.

During a call, the BS makes signal strength measurements of its own mobiles. Also, a receiver is used to monitor signal strength of users in other cells. Based on the measurements forwarded by the BS, the MSC makes decisions about call handoffs whenever needed.

A major drawback of first generation standards is their incompatibility, due, for example, to different channel spacing or frequency band of operation. This problem was especially severe in Europe, where many different countries had different standards, preventing users from roaming abroad. This is one of the reasons why the second generation pan-European GSM has been so successful both in the standardization process and commercially.

B. Second Generation Cellular Systems

The main difference between first and second generation cellular systems is that the modulation format used in the latter is digital rather than analog. This introduces considerable flexibility in handling the information signals, allowing the use of error control coding, source coding and compression, TDMA, and CDMA. Also, digital modulation is inherently more robust to channel impairments than frequency modulation, even though channel errors may result in artifacts which are less tolerated by human perception. Also, digital transmission increases the compatibility with the fixed portion of the telephone network (which is mostly digital), allows greater switching and routing flexibility, and lends itself to implementation in VLSI. Most importantly, the digital transmission format is more suitable for data communications, which is expected to be a major player in future systems and is already supported by some second generation systems. The coverage provided by second generation systems is not yet universal, but roaming is fully supported at least within Europe (this was a major goal of GSM), within North America, and within Japan.

There are basically two different types of second generation systems, depending on whether they are TDMA-based or CDMA-based. The first class includes standards such as IS-54/136* (the "digital AMPS") in the United States, GSM in Europe, and PDC in Japan. The IS-95 system is a CDMA-based standard, widely deployed in the United States and in Asia. All these systems use hybrid multiple access schemes. For example, in GSM each radio channel (with a data rate of 270 kbps) accommodates 8 users in a TDMA structure, but radio channels are divided through FDMA. Also, in IS-95 the basic radio channel is 1.25 MHz wide, so that 10 such channels in each direction can be supported in an FDMA arrangement for a total system bandwidth of 25 MHz.

1. TDMA-based Systems

In GSM, which is an example of a TDMA-based second generation system, the voice signal consists of a 13 kbps stream produced by the RPE speech coder. This bit stream is subdivided into two classes. The bits which are more critical in the speech reconstruction at the receiver are given protection and encoded with a rate 1/2 convolutional code, whereas the others are transmitted unprotected. Additional protection is provided by the interleaving mechanism which breaks the channel error correlation and makes the use of the error correction code more effective. This coding scheme introduces an additional 10 kbps to the original bit stream which, with the further addition of synchronization bits, training sequences for equalization and guard times, sums up to a total of about 33.8 kbps, which is the channel rate for each user (a 156-bit time slot every 4.6 msec).

In addition to the common control channels already mentioned (broadcast, paging and random access), GSM provides two additional channels, called associated control channels (ACCH). The slow ACCH (SACCH) is assigned as part of a connection, and is implemented by using dedicated frames in the GSM multiframe structure. The data rate on the SACCH sums up to a little less than 1 kbps. A typical use of the SACCH is the transmission of information about the signal strength received by the mobile from a number of surrounding BSs, to be used in the event that a handoff is needed. The fast ACCH is not continuously present, and is activated only under exceptional circumstances (e.g., while performing a handoff), when the transmission capacity normally assigned to the voice signal is preempted and reassigned to control traffic. Therefore, unlike in the analog systems, when a call is initiated a user is assigned a two-way control channel (SACCH) in addition to the two-way voice channel.

Mobile assisted handoff is implemented in GSM. Each cell has a broadcast channel which sends a pilot signal. Via its SACCH, a user learns the identities of the broadcast channels of the neighboring cells. During idle time, taking advantage of the TDMA structure where a user is active only one eighth of the time in either direction, each user listens to the broadcast channels of the closest 32 cells and reports data about the 6 strongest to its own BS via the SACCH. This information is periodically updated and used to determine candidate BS if the need for a handoff arises.

The various BS are interconnected through a backbone network, which also includes MSCs and their databases. In order to manage a complex network with a variety of features, GSM proponents have chosen to rely on some already defined procedures and interfaces, possibly with some adaptation for the mobile environment. GSM uses the CCITT signaling system number 7, originally defined for signaling in wireline ISDN.

An evolution of GSM is the DCS1800 system, whose architecture is very similar to what has been described so far. Differences with GSM are the frequency range (1800 rather than 900 MHz), the amount of spectrum allocated (150 rather than 50 MHz), the cell size (microcells rather than macrocells). Other TDMA-based standards include IS-54 and PDC. Similar to GSM, they use speech compression and channel coding, organize transmissions in frames, provide associated control channels, and support MAHO. Specific differences among the three systems are detailed in Table I.

*The significant differences of IS-136 over IS-54 are higher rate control channels and a richer set of services offered.

TABLE I Summary and Comparison of Second Generation TDMA-Based System Parameters

	Europe (GSM)	North America (IS-54/136)	Japan (PDC)
Access method	TDMA	TDMA	TDMA
Carrier spacing	200 kHz	30 kHz	25 kHz
Users per carrier	8 (16)	3 (6)	3
Modulation	GMSK	$\pi/4$-DQPSK	$\pi/4$-DQPSK
Voice codec	RPE 13 kbps	VSELP 7.95 kbps	VSELP 6.7 kbps
Voice frame	20 ms	20 ms	20 ms
Channel code	Convolutional	Convolutional	Convolutional
Coded bit rate	22.8 kbps	13 kbps	11.2 kbps
TDMA frame duration	4.6 ms	20 ms	20 ms
Interleaving	40 ms	27 ms	27 ms
ACCH	Extra slot	In slot	In slot
Handoff	MAHO	MAHO	MAHO

2. The IS-95 CDMA System

As already mentioned, IS-95 is based on CDMA, which takes advantage of spread-spectrum modulation in order to guarantee reliable communication in the presence of interference and multipath. spread-spectrum communication technique has its origins in military communications where antijamming (AJ) and low probability of intercept (LPI) capabilities are the main objectives. The information bearing signal (voice or low speed data), which is typically narrowband in nature (few tens of KHz bandwidth), is spread over a much wider bandwidth (typically of the order of several MHz) before transmission. The ratio of the transmission bandwidth (spread bandwidth) to the information bandwidth is known as the *processing gain*. The larger the processing gain, the better the system's immunity to jamming and intercept. Typical processing gains in the range 100 to 1000 are common.

Spectrum spreading is typically realized through one of two ways, namely, direct sequence (DS) or frequency hopped (FH) spread-spectrum (SS) technique. In DS-SS, the information signal is multiplied by a spreading signal which is generated at a much faster rate compared to the information rate. In FH-SS, narrowband modulation is used, but the instantaneous frequency of the modulating carrier is varied in a predetermined pattern. The spreading signal in DS and the hopping pattern in FH are generated using pseudo-random noise (PN) sequence generators. In order to correctly demodulate the information signal at the receiver, knowledge of the PN sequence used for spreading at the transmitter is needed. The receiver essentially correlates the received signal with a local replica of the PN signal (the operation is known as despreading), following which the data is demodulated.

Another important application of SS is *multiple access*, i.e., shared use of a common channel by several transmitters. Different transmitters can be assigned different spreading PN sequences (or spreading codes) and the central station can employ several correlation receivers each tuned to a given transmitter's spreading code. Such a multiple access system which uses DS-SS is known as the direct sequence code division multiple access (DS-CDMA). The IS-95 CDMA system has successfully exploited the multiple access capability of DS spread spectrum in cellular communications.

Since the users are separated in the code domain, CDMA users can transmit *simultaneously* in time over the entire available bandwidth. Also, due to the fact that the bandwidth of the transmitted signal is much larger than the bandwidth of the information signal (a factor of 128 is specified in IS-95), CDMA can enjoy the benefit of frequency diversity, since different spectral components can fade independently. In the time domain, this corresponds to being able to discriminate among the various replicas of the signal arriving at the receiver with different delays, which can therefore be constructively added rather than randomly interfere. The particular structure which is able to combine signals in this way is called RAKE receiver.

Another advantage of SS systems is the possibility of trading bandwidth expansion for coding redundancy. For example, if the channel bandwidth is 100 times larger than the user's signal bandwidth, it is possible to transmit such signal uncoded with a spreading factor of 100, or coded with a rate-$1/2$ code and a spreading factor of 50. This allows the use of coding (which is essential in certain applications) without further increase in the bandwidth of the transmitted signal. On the other hand, in TDMA-based systems the use of coding requires bandwidth expansion, unless trellis coded modulation (TCM) is used.

Due to the fact that all users share the same bandwidth and that interferers are essentially seen by the intended receiver as additive noise, any technique that reduces interference or enhances signal detection (i.e., reduces the E_b/N_o required to achieve a desired bit error rate) can result in an improvement of the CDMA system capacity. Also, this characteristic provides graceful performance degradation as users join the system, so that CDMA systems have "soft capacity," in the sense that a user can always be admitted to the system if a slight performance degradation for all users is accepted. This is not possible in static resource assignment systems (such as TDMA and FDMA with fixed channel allocation), where there is a constant number of physical channels per cell, and one more cannot be added. However, dynamic resource assignment has the potential to offer soft capacity even in these systems.

Typical techniques to enhance detection are the use of antenna diversity and of error correction coding (convolutional codes with rate $1/2$ on the base-to-mobile link and rate $1/3$ on the mobile-to-base link are specified in IS-95). Interleaving is also used to improve the error correction coding performance. Typical techniques to reduce interference are cell sectorization (120° antennas are used, which pick up only one-third as much interference as omnidirectional antennas) and silence suppression (during periods when the user is silent, which amount to more than half the time in typical telephone conversations, the transmitter switches to a low-rate mode used for signaling and synchronization purposes only).

A key requirement for DS-CDMA systems to work properly is that all signals reach the receiver with the same power, as power imbalances may greatly reduce the system capacity. To this purpose, tight power control must be used in the mobile-to-base direction, in order to compensate for the channel attenuation. If perfect power control were possible, all signals coming from mobiles in the same cell would have exactly the same power at the BS. On the other hand, power control errors do occur and interference from other cells cannot be controlled, and these effects degrade the performance of the system. In any event, accurate power control is a critical requirement of CDMA-based systems for proper operation, unlike in other systems where power control may increase the performance to some extent but is by no means critical. Open-loop power control is implemented based on the measurements of the pilot signal transmitted by the BS. Closed-loop power control is also specified in IS-95, with the BS sending one-bit power control commands every 1.25 ms, indicating whether the mobile should turn up or down the transmitted power in steps of 1 dB. A power dynamic range of 80 dB is typically required to track and compensate the received power variations due to fast multipath fading. In the other direction (base-to-mobile), a simpler power control scheme is used, whose main goals are improved transmission quality and minimized interference.

As in other systems, paging and random access channels are present. The broadcast channels are used by the BS to transmit a pilot signal, which is used by the mobiles as a power control reference and also as a coherent reference to be used for signal detection. A synchronization channel is also provided, which is used in the acquisition phase as a timing reference.

MAHO is specified in IS-95, where a mobile monitors the pilot signals coming from a number of cells and reports the results of these measurements to the controller. A unique feature of IS-95 is the so-called "soft handoff," where during the handoff procedure a mobile may be connected to more than one BS simultaneously. This allows the use of macrodiversity reception (the network can combine signals received by multiple BS), and provides a means to avoid interruptions as, unlike in other systems, the connection to the new BS is set up before the connection with the old BS is torn down (i.e., "make before break").

C. Cordless Systems

Another class of popular wireless devices is that of cordless systems. The first generation of such systems is very simple, consisting of a base which communicates with a single portable unit, and basically represents a wireless extension of the household telephone. Because of their convenience, such systems have been very successful, but they are subject to interference and lack of security. Also, with the emergence of the concept of personal communication services, the desire for digital devices and for increased capability (including some mobility) has been growing.

Second generation cordless systems have therefore been introduced, in which the cordless phone becomes part of a geographically distributed network and can connect to various attachment points, instead of being constrained to its own base equipment. For example, a service called Telepoint has been offered in the U.K., where a cordless phone can connect to the wireless equivalent of public telephone booths.

Probably the most popular of second generation systems is the Digital European (Enhanced) Cordless Telecommunication (DECT) system. It uses radio channels spaced by 1.728 MHz, each with a data rate of 1.152 Mbps and capable of supporting 12 TDMA channels. All carriers are used throughout the system (no frequency reuse plan), and channels are dynamically selected based on channel measurements at the time of connection setup. Speech coding uses ADPCM which does not offer good compression performance (32 kbps per voice channel) but

has very good robustness, especially needed in the very hostile environment (there is neither *a priori* separation between co-channel interferers nor SS protection). The two directions of communications share the same channel by time division duplexing (TDD). A DECT frame is made up of 24 slots, the first 12 carrying base-to-mobile traffic and the latter 12 carrying the mobile-to-base traffic. Control channels are provided by means of a dedicated field in each time slot.

Applications of DECT include residential systems (as with today's cordless phones), cordless PBXs with switching capability, telepoint public access (especially in high-density areas such as airports, train stations, and shopping and business centers), and radio in the local loop.

III. WIRELESS DATA SYSTEMS

With the increasing demand for data services, cellular telephony is not adequate to meet all the users' needs, and therefore provision of data services at increased data rates over wireless is gaining importance. In this section, we will discuss wireless data systems in the local as well as the wide-area scenarios. The newly emerging personal area networking (PAN) technologies are also discussed.

A. Wireless Local Area Networks

With the need for mobility and the positive experience with wireless telephony, it is natural for users to be attracted by wireless capabilities in the computer communications domain, so that the demand for wireless LANs has fueled a considerable amount of research and a number of commercial products.

Wireless LANs have many advantages over wired LANs, including quick and simple installation (no need for wiring), easy reconfigurability, application in environments where wiring is difficult or impractical (e.g., historical buildings), and support for ad hoc and temporary facilities (e.g., during a convention). Moreover, they can be used in environments where mobility is desired (e.g., inventory of stocks, car rental check-in).

The main differences between wireless LANs and cellular telephone systems are as follows. First and second generation cellular systems are targeted toward circuit-switched low-rate voice communications, provide wide-area coverage, support mobility of terminals, and have switching capabilities. On the other hand, wireless LANs are targeted toward computer and data applications, support high transmission speeds (up to several Mbps), are based on packet-switching technology, have limited switching capability and support no or limited mobility, and are intended as local area facilities. It is therefore not surprising that a completely separate set of technological solutions has been devised for cellular systems and wireless LANs (as happened for wireline telephone and computer networks). On the other hand, it must be said that the current trend toward integration of communications and computing will bring the two fields closer and closer together, and future wireless systems (addressed in a later section) will grow out of a combination of these technologies, as already testified by many proposals about third generation cellular systems which share many aspects of computer communications protocols.

The major difference between wireline and wireless LANs is in the data rates and quality they can support. Providing high transmission speeds in the wireless environment is very challenging because of the very hostile propagation environment. The IEEE 802.11 and HIPERLAN standards offer channel speeds of 1–2 Mbps and 25 Mbps, respectively (to be compared with 10 Mbps of Ethernet and 100 Mbps of fast Ethernet), and can operate over limited distances (tens of meters or less, much smaller than those allowed in wireline). Higher data rates can be achieved by means of Infrared communications (as opposed to radio transmission, implicitly considered so far), which, on the other hand, operates on shorter distances and usually requires line-of-sight propagation.

In an environment where data users are present and generate traffic in packet bursts, it is well known that circuit-switching and fixed resource assignment techniques (e.g., TDMA or FDMA) may be very inefficient. For this reason, other multiple access techniques have been devised, which are more suitable for this environment. The range of possible solutions has at one extreme complete coordination among users, and at the other extreme no coordination at all. An example of complete coordination may be polling systems, where a central controller tells users when they should transmit, based on the knowledge of their communications needs. This kind of structure may be more flexible than fixed resource assignment, but requires coordination among users and controller, and in particular some signaling is needed and some overhead is involved. On the other hand, the best example of uncoordinated access is the ALOHA protocol, in which users transmit whenever they have packets, regardless of what other users are doing. Overlapping transmissions (collisions) may lead to packet losses, which are recovered by retransmission.

The idea underlying the ALOHA protocol, and in particular its extreme simplicity, has been very successful, and has originated the whole research field of random access protocols.* Several modifications have been proposed to the original idea, trying to solve some problems that

*The ALOHA protocol has also great practical relevance by itself and is used in a variety of situations, including the random access channels of cellular systems.

random access presents. The first improvement was to slot the time axis, so that packets can be transmitted only at some fixed instants of time, thereby reducing the probability of overlapping transmissions. The recognition of the fact that the throughput performance greatly depends on the occurrence of collisions led to more advanced schemes, where transmitters listen to the channel before attempting transmission, and refrain from sending packets if the channel is sensed busy. In the presence of zero-signal propagation delay, this would lead to perfect sensing and would completely eliminate collisions. On the other hand, since signals take time to propagate, the sensing information may be incorrect. For example, just after a transmission is initiated a transmitter would sense the channel as idle and also start transmitting, thus causing a collision. Another improvement led to schemes where overlapping transmissions can be detected and aborted, thus minimizing the waste of communications resources caused by collisions. This idea is implemented in the carrier sense multiple access with collision detection (CSMA/CD) protocol, which is the scheme used in Ethernet. Other more elaborate improvements have also been proposed, addressing other more subtle issues such as protocol stability.

Given the large amount of research performed in the 1970s and 1980s on multiple access schemes for LANs, trying to extend those protocols as they are to the wireless domain seems a reasonable thing to do. However, this gives rise to problems which are not present in wireline LANs, and which effectively prevent from using the well-established LAN protocols in the wireless environment. One such typical issue is the "hidden terminal" problem. Basically, in wireline communications, all users are attached to a cable, and therefore they can all hear each other's signals. This condition is at the basis of the carrier-sensing protocols. Unfortunately, in radio communications this is not necessarily the case. In fact, the erratic behavior of the radio channel may not guarantee connectivity. Moreover, the fact that two mobile users A and B are within range of the same BS does not necessarily mean that they are within range of each other (e.g., if they are located at the extremes of a diameter of the BS's coverage region). Therefore, CSMA protocols as they are fail to work in the wireless environment, and other provisions must be made to guarantee proper operation. One possibility is the use of a busy tone, which is activated by the BS when some transmission is active, so that all mobiles which could potentially start communications to that BS (and are therefore within its range) will be informed about the channel being busy.* The multiple access protocol specified in CDPD (called DSMA/CD), which is the data services standard in the AMPS system, uses a busy/idle flag in the base-to-mobile direction to notify mobiles about the mobile-to-base channel occupancy, thus overcoming the hidden terminal problem.

In the two recent standards for wireless LANs, the IEEE 802.11 standard and the European HIPERLAN1, another protocol is used, which is called CSMA with collision avoidance (CSMA/CA). The basic concept relies on the observation that collisions are most likely to occur at the end of a transmission, since during that transmission communications requests may have accumulated and they are likely to be colliding as soon as the channel becomes idle. To avoid this effect, additional backoff mechanisms are introduced, in order to limit the number of terminals which will eventually attempt transmission.

In IEEE 802.11, the concept of interframe space (IFS) has been introduced. Interframe space is a time delay that a terminal has to wait before attempting transmission. Different lengths for IFS allow the introduction of different priorities. In particular, three IFS are specified, i.e., short IFS (SIFS) for high-priority control packets such as acknowledgements, point-coordination-function IFS (PIFS) for time-sensitive traffic, and distributed-coordination-function IFS (DIFS, the longest) for asyncronous traffic. If at the end of its IFS a terminal still senses the channel as idle (i.e., there is no higher priority traffic), it chooses a random time uniformly in a contention window (which is dynamically updated), and if after that time still no transmission is sensed, then the terminal sends its message. The hidden terminal problem can be solved by using some handshake procedure, where prior to message transmission there is a dialog between mobile terminal and BS, which improves reliability at the expense of some overhead.

More recent standards are based on technologies which can reach much higher data rates, of the order of tens of Mbps. IEEE 802.11a and HIPERLAN2 use the 5 GHz band, and are based on OFDM as a transmission technique. They achieve a maximum signaling rate of 54 Mbps at the physical layer, which provides about 32 Mbps of net user traffic. The main difference between the two standards is that IEEE 802.11a is based on connectionless transport, whereas in HIPERLAN2 a connection-oriented solution was preferred in order to more easily guarantee QoS and reserved bandwidth.

B. Personal Area Networks

Recently, wireless technologies for short-range, indoor communications among devices have been fast emerging with the aim to provide device mobility and system flexibility by replacing many proprietary cables that connect

*It should be noted here that the concept of "radio range" is not as straightforward (early models based on circular areas are rather gross oversimplifications), but is rather a probabilistic concept, and operation can still suffer from radio propagation effects.

devices to one another. Widely referred to as personal area networks, examples of such technologies include Bluetooth and HomeRF.

Bluetooth is a short-range ($\leq$10 m), low-power (1–100 mW) technology which operates in the unlicensed 2.4 GHz ISM band. Since microwave ovens, WLANs, etc., can operate in the same band, in order to avoid interference to/from such sources, Bluetooth system has been designed based on FH-SS and TDD. The system uses 79 carrier frequencies over 1 MHz bandwidth. The nominal bit rate of transmission is 1 Mbps. Bluetooth allows a collection of devices to form small overlapping networks (known as 'piconets') in an adhoc fashion. Devices such as PDAs, cellphones, laptops, desktops, etc., can be part of a Bluetooth system.

The unit that establishes and coordinates a piconet is called the "master." All other participants are called the "slaves." There can be a maximum of seven slaves in a piconet. Time is divided into slots of 625 μs. The timing is maintained by the master and slaves synchronize to the master's clock. A FH channel is associated with every piconet and each piconet has a unique hopping sequence determined by the master. The frequency hopping rate is 1600 hops/sec. A collection of piconets is called a "scatternet."

Higher layer packets are fragmented into Bluetooth physical layer packets. Each packet has 126 bits of control information and a maximum payload of 2745 bits. Packet transmissions on one hop frequency can take multiple slots. Full duplex communication is realized through TDD. Access control is exercised by the master. Only communication between the master and a slave is possible (i.e., slave-to-slave communication is routed via the master). The master transmits in the odd-numbered slot and the corresponding slave responds in the next slot.

It is likely that the Bluetooth specification will be adopted in the IEEE 802.15 Wireless Personal Area Network standard.

HomeRF is another short-range, low-power wireless technology which will enable a class of mobile consumer devices drawing from the content of the Internet and home PC. HomeRF technology is supported by an industry association of several computer, telecom, and consumer electronics companies. The origins of HomeRF are from the IEEE 802.11 WLAN standard and the DECT system. IEEE 802.11 uses CSMA/CA for multiple access and is good for data. DECT, on the other hand, uses TDMA for multiple access and is good for voice. HomeRF essentially attempts to combine the best features of both IEEE 802.11 and DECT in order to provide data and voice communications through mobile consumer devices. Accordingly, the Shared Wireless Access Protocol (SWAP) has been developed as an open standard for voice and data networking. SWAP handles voice like DECT but using frequency hopping. SWAP handles data like IEEE 802.11 but uses relaxed PHY layer specifications to reduce the cost of the mobile device.

C. Wireless Data in Cellular Systems

The first and second generation (1G and 2G) cellular systems described before provide primarily voice services to mobile subscribers and to some extent data services at low data rates. Data services on 1G and 2G systems are essentially provided on circuit-switched air links. That is, an over-the-air traffic channel is brought up and held through the entire duration of the data call. No changes are needed to the physical layer to provide data services in this case. However, suitable protocol stacks need to be provided at the mobile station, base station, and the MSC in order to enable data services. At call origination, the service option parameter can specify a certain data service instead of voice service.

In 1993, the United States cellular industry developed the Cellular Digital Packet Data standard (CDPD), intended to coexist with the 1G and 2G cellular systems in the United States, and to use the 30 kHz radio channels on a shared basis to provide packet-switched data communications. The CDPD standard takes advantage of the fact that some radio channels in cellular systems are often left unused during normal operation,* and in this case the corresponding radio resources can be devoted to data transmission. By doing so, CDPD is able to provide packet data services without additional bandwidth. CDPD uses the existing cellular BSs and air interface, but not the MSCs, as it has its own routing capabilities.

CDPD uses the 30 kHz AMPS radio channels to provide 19.2 kbps data. The use of the cellular radio channels is only allowed when such channels are not supporting telephone calls, which have preemptive priority over the data traffic. In the standard specifications, when a channel used by CDPD is assigned to a telephone call, another empty channel is sought, so that the CDPD connections "hop" among various channels according to the channel assignment dynamics of the cellular network. In practice, however, in many cases a voice channel per cell is permanently reserved for CDPD use, as this is much simpler to manage (although less efficient).

The multiple access protocol in CDPD is DSMA/CD, which belongs to the class of carrier-sense random access schemes described in the previous section. Data-link functions are provided by the Mobile Data Link Protocol, whereas the Radio Resource Management Protocol

*Note, in fact, that the number of available channels must be chosen to guarantee adequate blocking performance under peak conditions.

is responsible for finding and managing unused cellular radio channels. Services provided by CDPD include broadcast, e-mail, monitoring, and dispatch. Error control capability is also provided, through the use of a Reed Solomon code.

In the GSM system, data services including asynchronous data (computer-to-computer traffic) and digital fax are supported at 9.6 Kbps/14.4 Kbps rate. Protocol stacks are provided at the mobile station, BS, and the MSC in order to facilitate the emulation of the line/fax modem functionalities over-the-air. Transparent (only FEC is used to ensure link reliability) and nontransparent (both FEC and link layer ARQ are used) modes of these data services are possible. Increased data rates up to 115 kbps can be achieved through high-speed circuit-switched data (HSCSD) services in GSM. The main idea in HSCSD is to assign multiple time slots per TDMA frame to subscribers who need high data rates. Up to 8 slots ($8 \times 14.4 = 115.2$ kbps) per TDMA frame can, in principle, be assigned to each HSCSD subscriber.

The IS-95A CDMA system too supports data services including (1) asynchronous data (modem emulation), (2) digital fax, and (3) packet data. Here again, packet data service for e-mail and web browsing is essentially "packets over a circuit switched air-link." The maximum data rate supported is 9.6 kbps. IS-95 CDMA employs a transmission control protocol/radio Link protocol (TCP/IP/PPP/RLP/IS95) based stack at the mobile and the BS to provide reliable data services. The RLP performs a partial link layer error recovery through a limited number of RLP frame retransmissions in case of frame error. Specifically, RLP uses a negative acknowledgement (NACK) based selective repeat ARQ to improve the frame error rate seen by the TCP layer. The performance of the protocol stack is mostly determined by the interaction between the TCP and RLP layers under varying channel conditions. The effective data rate seen by the user is typically in the range of 7–8 kbps depending on the channel conditions and excluding the protocol overhead.

An interesting feature in IS-95 packet data service is the dormant mode. In dormant mode, if the user does not send/receive data for a specified amount of time (this can happen, for example, when reading the contents of a web page after downloading it), the system takes away the traffic channel resources at the physical layer without disturbing the higher protocol layer (PPP) states. Traffic channel is again established once packets start to flow again. This enables traffic channels to be allotted mainly to data users with active data traffic exchange over-the-air. Keeping the PPP states helps to achieve faster reconnection (since the PPP link need not be renegotiated during reactivation). Dormant mode results in efficient use of traffic channels, thus increasing the revenue of the service provider while saving air-time charges for the subscriber when idle.

Higher data rates in IS-95 CDMA are made possible through "code aggregation" (similar to slot aggregation in HSCSD in GSM). That is, multiple Walsh codes can be assigned to mobiles requiring higher rates. In such a system (known as IS-95B), up to 8 traffic channels can be aggregated giving a maximum of $8 \times 14.4 = 115.2$ kbps to the users. The system also supports *simultaneous* transfer of voice and data from one mobile station.

It is well known that computer-generated data traffic is bursty in nature and circuit switched connections (as provided by 1G and 2G systems) to transport bursty data traffic is inefficient. Packet-switching is more efficient for transporting bursty traffic. With the increasing demand on mobile data services, the evolution of 2G systems to next generation (3G) systems has been focused on supporting "packet-switched operation over-the-air" in order to make more efficient use of the radio resources.

1. GPRS

General packet radio service (GPRS) is a "packet mode" wireless data system that has been standardized to operate on GSM infrastructure, by introducing new packet support nodes and associated protocol stacks. A portion of the radio resources (channel frequencies) in an existing GSM system may be dedicated for packet data services using GPRS. Alternatively, GPRS and GSM services may dynamically share the same radio resources. Thus, GSM voice services and GPRS packet data services can coexist in the same GSM system.

GPRS provides IP connectivity to the mobile users. The maximum data rate supported is 171 kbps realized through statistical multiplexing of traffic over-the-air. Point-to-point (e.g., e-mail, Internet access, remote access, road toll) as well as point-to-multipoint (e.g., multicast, financial updates, fleet management, traffic information) applications are envisaged.

The new packet support nodes in GPRS are Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN). The SGSN and GGSN enable the mobile data users to connect to the Internet, essentially playing analogous roles of the Serving MSC and Gateway MSC in GSM that connect voice users to the PSTN.

The over-the-air communication between a mobile station (MS) and the GPRS network is defined by the physical layer and the data link layer. The physical layer functions involve modulation, demodulation, channel encoding/decoding, etc. The data link layer consists of two sublayers, namely, logical link control (LLC) layer and the radio link control/media access control (RLC/MAC) layer.

The LLC layer operates between the MS and the SGSN, and provides a logical link between them. The LLC layer supports two modes of operation, namely, asynchronous disconnected mode (ADM) and asynchronous balanced mode (ABM). In ADM, the LLC performs "unacknowledged" operation and does not provide any error recovery procedure to guarantee in-order deliver (such a mode may be applicable for delay-sensitive, error-tolerant applications like voice over IP). In ADM, the LLC performs "acknowledged" operation which includes functionalities like error recovery through ARQ, in-order delivery, and flow control (these functionalities are essentially based on the well-known HDLC protocol). Also, LLC ensures data confidentiality through ciphering functions.

The RLC/MAC layers are primarily responsible for the efficient sharing of common radio resources by several mobiles. The RLC/MAC peers are at the MSC and the BS. Each LLC frame is segmented into several RLC data blocks of fixed size. Each RLC data block occupies a fixed number of slots, depending on the channel coding scheme used. The RLC layer can operate either in acknowledged or unacknoledged mode. In acknowledged mode, RLC provides for the selective retransmission of erroneous RLC data blocks. The MAC operates on a slotted-ALOHA based reservation protocol. The MAC layer requests/reserves resources in terms of number of data slots. The MAC function provides arbitration between multiple mobiles attempting to transmit simultaneously, and provides collision detection and recovery procedures.

2. EDGE

Enhanced data rates for global evolution (EDGE) technology comes next to GPRS in the evolution path from GSM. EDGE supports packet services at data rates as high as 384 kbps. Such high rates have been made possible by using advanced physical layer design features including higher-order modulation (8-PSK) and link adaptation. The modulation and coding formats are changed based on the channel conditions. While other 2G systems achieve spectral efficiencies in the range of 0.05–0.15 bits/sec/Hz/cell, EDGE offers increased spectral efficiency of about 0.45 bits/sec/Hz/cell. The carrier spacing (200 KHz) and burst format (116 data symbols, 26 symbols in the middle for training, 3 tail symbols at either end, and 8.25 guard symbols at one end) are kept to be the same as in GSM so that reuse of and compatibility to existing GSM/GPRS infrastructure is easily achieved.

D. WAP and iMode

Internet protocols have been designed for wireline, high bandwidth, low latency, reliable links, and the Internet content is mainly designed for large-and high-resolution displays. But the wireless medium is inherently low bandwidth and error-prone, where Internet protocols like TCP perform poorly. Also displays in mobile devices are small in size and low in resolution. HTML, which is popular and good for large displays, is a poor markup language for small screens with low resolution. Unwired Planet (now Phone.com) first demonstrated a microweb browser for cellular phones in 1995. It tailored the HTML technologies for small screens and limited resources, and called it Handheld Device Markup Language (HDML). In 1997, Unwired Planet, Ericsson, Motorola, and Nokia formed the Wireless Access Protocol (WAP) Forum (since then several companies have joined the WAP Forum), to work toward a global, license-free standard based on the Internet protocol model optimized for wireless networks/devices. HDML and other pieces of its architecture eventually have become the Wireless Markup Language (WML) for authoring services. WAP can be used over a broad range of cellular data bearers, including GPRS, EDGE, and 3G cellular.

The WAP protocol stack is optimized for wireless. It consists of five layers (similar to OSI model) to provide flexibility and scalability. The various layers are application layer (WAE, wireless application environment), session layer (WSP, wireless session protocol), transaction layer (WTP, wireless transaction protocol), security layer (WTLS, wireless transport layer security), and transport layer (WDP, wireless datagram protocol). A key element in the WAP architecture is the WAP gateway which does the translation of protocols (from WAP to IP and vice versa) so that the limitations of the wireless networks/devices are transparent to the Internet. Specifically the Gateway translates the WAP protocols (WML, WSP, WTP, WTLS, WDP) to their corresponding Internet counterparts (HTML, HTTP, TLS, TCP/IP or UDP/IP).

iMode is a competing technology to WAP. Having its origin from NTT, Japan, iMode is a packet-switched service at 9.6 kbps rate meant to provide mobile access to specially formatted web sites and portals and let mobile users send/receive instant messages, e-mails, and photos taken with the phones. iMode has millions of subscribers in Japan with thousands of formatted web sites.

IV. THIRD GENERATION (3G) CELLULAR

Third generation cellular systems—popularly called 3G systems—are digital mobile systems having greatly enhanced performance and service characteristics including multimedia services (voice, data, high quality image, video, etc.) at high data rates (up to 2 Mbps) with different quality of service requirements. While the 1G

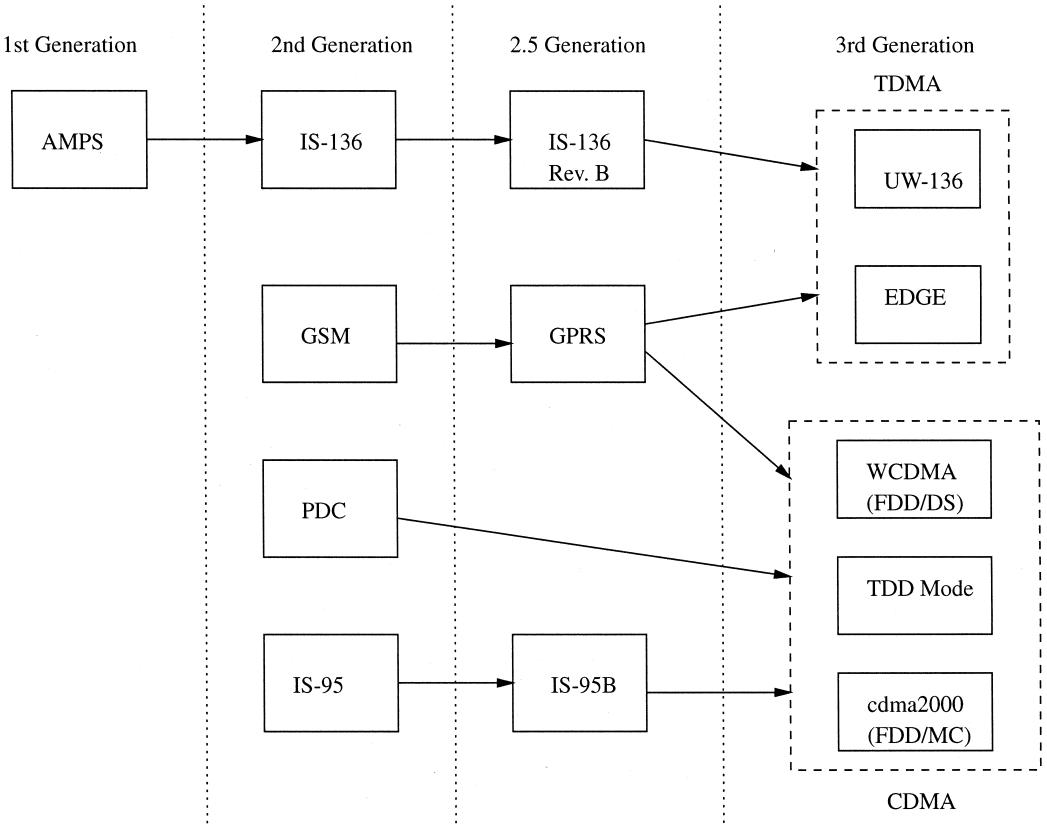


FIGURE 3 Evolution of cellular systems standards.

and 2G standards have been developed primarily by the regional and national standards bodies (e.g., ETSI, ANSI, TIA), the standardization efforts of 3G have been at the international level, keeping the evolving global market place in mind. The development of 3G standards/specifications were guided by the global requirement developed by the ITU which are referred to as IMT-2000. The evolution path of cellular system standards from first to third generation is illustrated in Fig. 3. The vision of 3G technologies has been based on wireless services that integrate multimedia, packet switching, and wideband radio with higher spectral efficiencies than the existing 2G technologies. In addition, the 3G standards would permit global roaming procedures with simple terminals or devices.

IMT-2000 specifies the minimum performance requirements to support packet and circuit data services at symmetric and asymmetric data rates in all environments —144 kbps in rural/outdoor vehicular environment (up to 250 km/hr), 384 kbps in urban/suburban outdoor environment (up to 150 Km/hr and 2 Mbps in indoor/local environment (up to 10 km/hr). Bit error rate and maximum delay performance for both real-time/constant delay and nonreal time/variable delay applications have been specified.

Based on several discussions toward harmonization of competing standards proposals, ITU has recommended 3 modes of CDMA-based 3G standards, namely direct sequence (DS) mode, multicarrier (MC) mode, and TDD mode. The DS mode will be based on WCDMA FDD at a chip rate of 3.84 Mcps, and the TDD mode will be based on UTRA TDD from 3GPP and TD-SCDMA from China. The MC mode will be based on cdma2000 at a chip rate of 3.6864 Mcps. While the MC mode would employ synchronous BS. DS and TDD modes would use asynchronous BS. All modes, however, would have to work with both dominant core network standards, namely GSM MAP and ANSI-41 based core networks, to enable interoperability across different networks.

Several countries (in Europe and most of Asia) will make available the IMT-2000 bands of 2×60 MHz (1920–1980 MHz) and (2110–2170 MHz) for WCDMA FDD (DS mode). The cdma2000 system (MC mode) is essentially a natural evolution of IS-95A CDMA and hence uses the same PCS spectrum allocation while maintaining backward compatibility to IS-95 systems.

Both WCDMA- and cdma2000-based 3G systems offer some unique physical layer features that were not present in 2G systems. These include: use of turbo codes, coherent

detection on reverse link, fast power control on the forward link, variable spreading factor to increase user data rate, use of transmit diversity on the forward link, use of smart antenna technologies, multi-user detection, and provision of varying QoS through a specialized LLC/MAC layer protocols.

V. RESEARCH AND NEW TECHNOLOGIES FOR 3G AND BEYOND

Future wireless networks are expected to be deployed as a further evolution of 3G and personal communications technologies. Even though the phrase "fourth generation systems" may still mean different things to different people, there seems to be general consensus about some of the key requirements of these new systems, which include global coverage, high capacity, improved quality, and increased variety of offered services (including data, video, and multimedia). The distinction among various services in existence today (paging, cordless, cellular) will blurr and all of them are expected to be offered by a single personal device. The broadband ISDN will be used to provide access to information networks and servers (e.g., the Internet), and satellites will be part of this picture as well. Applications including multimedia sessions are envisioned, with telephone conversations accounting for the minority of the total traffic. Applications which are already a reality in the wireline domain, such as web browsing, multimedia communications, teleconferencing, intelligent networking, virtual offices, and so on, will find their way into the wireless domain, allowing people to be set free from location constraints.

A great deal of research has been and is being done on all aspects of third and fourth generation wireless systems. In the following subsections, we give a quick overview of the many issues involved and of some prominent new technologies which have come to be considered key enablers of the third/fourth generation wireless vision. Details about any of them can be found in the technical literature (see Section VI for a basic set of references).

A. Physical Layer and Error Control

Advanced requirements envisioned in future wireless networks include data rates which, if not quantitatively comparable with those of the broadband ISDN, at least qualitatively are broadband and intended to support multimedia. This often leads to specifications of channel data rates in the range of 20 Mbps or higher. Such data rates are hard to achieve on a radio channel, especially because of the need for equalization due to severe time dispersion of the transmitted pulses. Considerable overhead may be involved when using standard equalization techniques. Use of advanced signal processing may be needed, including antenna array adaptation and combining which is being considered as a feasible means to counteract multipath and interference.

As to the carrier frequency range to be used, current systems occupy the 900 MHz band and 1.8–1.9 GHz band, whereas new allocations are being made at higher frequencies (see Ayanoglu *et al.*, 1996 for a chart with specific information about spectrum allocations). Issues which arise when choosing the operating frequency include regulatory policies (some bands do not require licensing and are therefore attractive both for the small cost and for the immediate availability, whereas others require licensing and FCC approval, which is a long and expensive process but guarantees dedicated channel use), cost of the equipment (which is higher at higher frequencies in general), propagation characteristics (e.g., at 60 GHz electromagnetic signals do not propagate through walls, and this may be an advantage for indoor systems due to reduced interference but also a serious drawback for global coverage), and amount of bandwidth available (usually, the higher the carrier frequency, the larger the available bandwidth). The use of diffuse infrared propagation does not seem to be a good alternative to radio systems due to the effect of multipath and the very limited range (line of sight is typically required), although it may be preferable when concerned with electromagnetic interference (e.g., in hospitals).

Linear modulation techniques can be used in wireless systems, e.g., BPSK, QPSK (with their differential counterparts), and QAM, even though demodulation of multilevel signaling is usually hard to implement in fading environments. Schemes such as GMSK or $\pi/4$-DQPSK achieve (or approximate) constant envelope characteristics, thus enjoying the advantage of better efficiency of the amplifiers (which need not be linear). Much attention is being given to orthogonal frequency division multiplexing (OFDM), a technique which uses many narrowband carriers instead of a single broadband channel, thus simplifying considerably the need for equalization (Bingham, 1990). Multicarrier modulation techniques are also attractive for multimedia traffic, as they enjoy additional flexibility as to variable data rates, and therefore can easily support bandwidth-upon-demand. The use of SS modulation is also being considered with great interest, due to its inherent ability to deal with multipath propagation and co-channel interference, especially in its wideband version. Multicarrier CDMA techniques are being investigated with an intention to exploit the OFDM as well as CDMA advantages.

Regarding error control, two basic techniques exist, namely forward error correction (FEC) and automatic

repeat request (ARQ) (Lin and Costello, 1983). In FEC schemes, redundant symbols are added at the transmitter in such a way that the receiver is able to correctly decode the message even in the presence of transmission errors (to a certain extent). These techniques provide deterministic transmission delay,* but usually involve considerable overhead. In ARQ schemes, on the other hand, only few redundant symbols are added before transmission, giving the receiver the capability to detect errors and to request retransmission of the message portions which get corrupted on the channel. A much smaller overhead is paid in this case, but no guarantees can be made about transmission delays, as a packet may need to be retransmitted many times before it is finally received correctly. Also, ARQ systems require a feedback channel which is not needed for FEC. For future services, where small radio cells will lead to very short propagation delays and will allow almost immediate feedback, ARQ techniques may gain renewed interest. Studies have been recently published, showing that for packet communications in the local wireless environment ARQ schemes can successfully compete with FEC, due to bandwidth efficiency and additional features such as the possibility of selective packet dropping in delay-constrained communications, easier QoS support, enhanced flexibility, and lower complexity (Khansari *et al.*, 1996 and Zorzi and Rao, 1997a). It is likely that hybrid schemes (possibly adaptive) will be employed in these systems, according to the trade-offs between data reliability on the one hand and complexity and delay performance on the other. More advanced techniques, such as TCM, (Bigliari *et al.*, 1991) and Turbo Codes (Berrou *et al.*, 1993) which provide much better protection against interference and noise, suffer from other problems (e.g., decoding complexity for TCM and interleaving delay for Turbo Codes), which might delay their application to broadband wireless communications.

1. Turbo Codes

Channel coding provides improved error performance in digital communication systems by mapping input sequences to code sequences which add redundancy to the transmission. It is well known that the channel capacity (i.e., the maximum rate at which reliable communications is possible) is approachable by random codes as the length of the transmitted data block is increased. However, as the block length increases the decoder complexity becomes unmanageable. Therefore, it has been a goal to design codes which perform near the capacity while offering modest encoding and decoding complexity.

*This delay is due to the encoding/decoding algorithms and to the use of interleaving for improved error correction performance.

In 1993, a very powerful channel coding scheme was developed by Berrou *et al.* (1993) which uses the ideas related to block and trellis codes. This scheme, which the inventors have named as *turbo codes*, uses simple convolutional codes separated by interleaver(s) to produce generally low rate block codes. The decoding of turbo codes is performed by decoding the individual convolutional codes separately and sharing bit reliability information in an iterative manner. This results in a block coding technique which mimics the performance of random codes but whose decoding complexity is only loosely tied to the block size. Furthermore, it has been shown that turbo codes are capable of achieving near capacity performance for sufficiently large block lengths. In particular, on AWGN channels, turbo codes have been shown to provide 10^{-5} BER at an E_b/N_o of 0.7 dB. It is to be noted that (1) an uncoded BPSK scheme would require 9.6 dB E_b/N_o to achieve the same BER, and (2) the theoretical AWGN channel capacity (due to Shannon) is -1.6 dB E_b/N_o.

Turbo codes are used in 3G cellular systems, and are expected to be widely used in the future wireless systems as well. The relevance of using turbo codes in 3G systems is as follows. In 3G, CDMA has been the choice for multiple access. Inherently, CDMA systems are interference limited (Viterbi, 1995). In other words, any technique which could reduce the E_b/N_o required to achieve a desired BER performance of the link can increase the system capacity in CDMA. Use of turbo codes provides an effective means to reduce the required E_b/N_o and hence increase the CDMA system capacity. In cdma2000, the supplementary channels on forward and reverse links (F-SCH and R-SCH) optionally use turbo codes. In WCDMA, turbo codes are optionally used on the DCH, CPCH, DSCH, FACH.

2. Multiuser Detection

Another interesting feature in 3G is the optional use of multiuser detection receivers (Verdu, 1998) for improved performance in CDMA. The conventional CDMA receiver is a single user spreading code matched filter (MF) receiver where we treat the interference from other users as noise and detect only the desired user's signal in the presence of multiple access inteference (MAI). Multiuser receivers, on the other hand, will take into account the other users' data to demodulate the desired user. Hence multiuser receivers will need the knowledge of the other users' spreading codes, delays, phases, received powers, etc. The optimum multiuser receiver was derived by Verdu (1998), who also showed that the implementation complexity increases exponentially with the number of active users. Consequently, several recent research investigations have focused on low complexity, suboptimal multiuser detectors whose performance is better than

the conventional MF receiver (Xie, 1990 and Giannakis, 2000).

3. Transmit Diversity and Beamforming

As mentioned earlier, in mobile radio channels, diversity reception techniques are widely used to mitigate the effect of fading. There are different ways by which diversity reception can be achieved, namely, time diversity, frequency diversity, space diversity, multipath diversity, etc. In almost all cellular systems, space diversity is typically achieved by way of using two receive antennas at the BS (sufficiently separated in space so that the fades seen on each antenna are independent). Different ways of combining the faded signals through different diversity paths are possible. These include maximal ratio combining (MRC), equal gain combining (EGC), and selection combining (SC). Also, in DS-CDMA systems (e.g., IS-95A, 3G), multipath diversity can be exploited for improved performance through the use of RAKE receivers (Viterbi, 1995).

In addition to employing receiver diversity, 3G systems will employ transmit diversity as well. That is, the BS can employ multiple transmit antennas and send the transmit signal through all these antennas. Such channels/systems, widely known as multiple input multiple output (MIMO) channels/systems, have recently drawn considerable research attention owing to their potential for improved performance (Foschini and Gans, 1998). Also, error control coding for such MIMO channels, popularly known as spacetime codes, has evolved as an important research area in the recent years (Tarokh *et al.*, 1998).

In WCDMA, two types of transmit diversity schemes based on open- and closed-loop techniques are employed on the downlink (BS to mobile link). In cdma2000, transmit antenna diversity can be implemented on a multicarrier downlink, where a subset of carriers is transmitted on each antenna. The frequency filtering on different carriers provides near-perfect orthogonality between antennas resulting in improved capacity.

Antenna beamforming is another interesting feature in future wireless systems. Spot beams can be used in 3G to improve coverage, capacity, and data rate. In addition to existing sectors, embedded sectors toward "holes" (spots with weak/no coverage within a sector) can be provided by directing narrow, strong beams toward such holes to improve coverage. Additional beams could also be directed toward "hot spots" (high traffic generating area, e.g., mall, fair, etc.) within a sector so that increased capacity can be provided toward such hot spots. Further, a disadvantaged user (say, located at the edge of the cell) demanding higher data rates may be satisfactorily served through beamforming. It is noted that 3G systems enable such beamforming through the provision of "auxiliary pilot signals" in addition to the normal pilots. 3G systems will support detecting auxiliary pilots and soft handover of a mobile entering/leaving a spot beam through normal handoff procedures.

B. Multiple Access

In cellular systems, two issues must be solved at the multiple access level. The *channel reuse* scheme determines how channels are allocated to cells throughout the system, whereas the *access scheme* solves the problem of resource sharing within a radio cell. Different systems make different choices on these two issues. For example, classic cellular systems use some frequency reuse plan, so that the channel reuse leads to tolerable co-channel interference. Also, in circuit-switched systems, the access scheme is usually achieved by orthogonal channels (in the time or frequency domain). These channels can be rigidly assigned to cells (as in first and second generation systems), or more flexible schemes can be used, ranging from borrowing schemes to fully dynamic channel allocation. Various schemes have been proposed, achieving better resource utilization efficiency at the expense of additional complexity and the need for coordination among network elements (Katzela and Naghshineh, 1996). One attractive feature of dynamic resource allocation is soft capacity, which provides additional flexibility in traffic management and improved teletraffic performance (Grandhi *et al.*, 1997). In cellular CDMA, SS is used to achieve both multiple access and channel reuse. The inherent interference rejection capability of this technique makes it possible to use the same radio channel in all cells, thus avoiding the need for a frequency reuse plan.

Recent proposals for multiple access schemes make the implicit assumption that information is carried in packets. The flexibility of packet communications and the increased delay-tolerance of some data traffic opens up opportunities for more efficient design of multiple access protocols. Many of these protocols resemble those used successfully in local area networks, with modifications to make them suitable to the wireless environment. An early proposal, packet reservation multiple access (PRMA) (Goodman *et al.*, 1989), is essentially a reservation ALOHA scheme, where slots in a frame (the "channels") can be either reserved or available. Multipacket messages are transmitted on a contention basis in available slots, and a successful transmission acts as a reservation, guaranteeing contention-free access (in reserved slots) during the rest of the message. More elaborate access techniques have been proposed (Bianchi *et al.*, 1997; Karol *et al.*, 1995; Mitrou *et al.*, 1990; and Wu *et al.*, 1994).

The basic structure of most proposed multiple access protocols is based on reservation. It is assumed that some provision is given to the users to make reservations (as soon as they become active), thereby notifying the central BS of their need for communications resources. This can be seen as a separate channel, even though it can be dynamically implemented based on the instantaneous channel load (as in the case of PRMA, where increased traffic reduces the number of available slots). Information packets, on the other hand, are carried based on reservation and polling: users who gained access to the system (by successfully placing a reservation) are polled by the BS, according to some service policy which in general depends on the QoS requirements of each connection. Flexible protocols, capable of managing different traffic types and QoS requirements can be envisioned in this environment. Examples of these protocols can be found in (Bianchi *et al.*, 1997; Karol *et al.*, 1995; and Mitrou *et al.*, 1990).

Other interesting issues regard the compatibility of multiple access schemes with advanced physical layer techniques. For example, the use of array antennas may require the users to transmit pilot signals periodically or on demand, so that the BS can estimate the channel impulse response and determine the optimum weighting coefficients to be used in the diversity combining block. This requirement must be integrated in the design of the multiple access scheme (Acampora, 1997).

C. Quality of Service Support

With the advent of broadband networking and ATM, quality of service (QoS) provisioning has become a key concept in telecommunications (Schwartz, 1996). Typical QoS metrics in ATM networks are the packet loss rate, delay, and delay jitter (Onvural, 1995). Packet networks, unlike circuit-switched networks, are based on statistical multiplexing and resource sharing, and are prone to congestion, which occurs when the instantaneous users' demand collectively exceeds the network's capabilities. A major concern in broadband networking is to try to avoid congestion situations as much as possible and to cope with them when they occur. Admission control techniques are used in order to determine whether or not incoming connection requests can be accepted (Acampora, 1994a). A connection request can only be accepted if its QoS requirements can be met and if the QoS requirements of all connections already admitted can be maintained. This concept of admitting users based on a QoS calculation at the time the connection request is presented to the network implicitly relies on a "static" network paradigm, in which the topology of the network, the capacity of its links, and the geographic distribution of the admitted traffic are fixed. In this situation, if the admission control algorithm indicates that QoS is guaranteed when a connection is set up, then those guarantees are in general not violated in the future.

It is clear that this paradigm does not apply to wireless networks, where neither the network topology and the capacity of its links, nor the geographic distribution of the admitted traffic, are fixed. In fact, link capacity fluctuates due to fading and mobility, and sometimes links become unavailable. Also, mobility is a primary requirement in wireless communications, and users can roam anywhere after being admitted, possibly driving some radio cells to overload. In this environment, guaranteeing QoS is an exceedingly harder task, if at all possible. On the other hand, QoS is an essential feature of modern telecommunications services, so a way must be found to offer this capability in the wireless domain as well (Naghshineh and Acampora, 1996).

Of course, a very strict design of the network architecture and of its protocols, where everything is overdesigned in such a way that worst-case situations are still manageable, is a legitimate solution to that problem (using essentially the same philosophy as circuitswitching). However, its inefficiency makes it virtually useless for advanced wireless systems. A better solution may be to accept a different definition of QoS guarantees, where a user negotiates a *set* of QoS levels, and the network tries to provide the highest possible at any given time (and bills the user accordingly). For example, if the bandwidth of a link drops because of degraded propagation conditions, the quality is downgraded and a lower service is provided. This may sound like a contradiction of the concept of QoS guarantees, but is to be compared with the dropping of the connection altogether due to the inability to maintain the required QoS level, a solution which in many cases may be less desirable.

An even more elaborate solution can be found in cooperative QoS management strategies involving both the application and the network (Naghshineh and Willebeek-LeMair, 1997). The network in general has requested to be application-aware, in the sense that it knows the application's needs (as specified in the QoS profile) and tries to meet them. The idea here is to also make the application aware of what happens in the network, giving it a chance to react and adapt to changed conditions. For example, this has the advantage that the application itself may choose how to modify the level of service, instead of letting the network (which in most cases does not have enough knowledge, e.g., of the application contents) do it. A typical example is layered encoding, as used, for example, in video streams, in which bits (packets) are divided into classes of different priorities. Highest priority information is essential for any understanding of the message, whereas a lot of low-priority information, although desirable for high-quality playback, can be lost while still

retaining sufficient understanding. Techniques like this can be used to dynamically adapt the perceived quality to the channel conditions. As long as the user/application is willing to tolerate these fluctuations, such techniques may provide a dramatic improvement in connection dropping performance as caused by radio channel impairments.

D. Networking and Transport Issues

Much research has been devoted to the analysis of higher layer protocols and to the design of schemes which can deal with the unique issues arising in the wireless environment.

One of the characterizing aspects of wireless networks is mobility, and this requires a whole set of new functionalities which need to be provided. Current cellular systems keep databases and offer handoff capabilities, which allow users to freely roam throughout the system area. Performing a handoff is not a major undertaking in current systems, due to traffic uniformity and low data rates. With broadband networking capabilities and QoS support, handoff will be much more challenging for a number of reasons. First of all, a brief interruption of speech communications, as occurs in hard handoff in current cellular systems, may be barely noticeable and is certainly accepted in general. For a broadband data connection, this interruption may involve a large amount of data which, in the case of reliable connections, must be retransmitted or recovered. Error recovery, as well as duplication and reordering problems, must be addressed in this case. Furthermore, in current systems handoff requests are usually accepted or rejected based on channel occupancy in the target radio cell, since cellular telephone networks have a very primitive concept of QoS, if any. On the other hand, when heterogeneous traffic classes are to be served, the QoS calculations to be performed by the admission controller are much more complicated and time-consuming. If these calculations are required every time a handoff is requested, the burden on the controllers may be exceedingly large, possibly causing a dramatic increase of signaling traffic and of the number of calls which are dropped because a handoff cannot be completed within reasonable time. Three types of handoff schemes exist (Wu and Chang, 1996): anchor rerouting, dynamic rerouting, and connection tree routing. Unlike the first two, the third class of schemes does not require tearing down and setting up the connection, and has been actively studied for broadband wireless applications (Acampora and Naghshineh, 1994b and Veeraraghavan *et al.*, 1997).

Other issues at the networking and higher layers have to do with the performance of established network protocols over mobile and wireless systems. A very good example is the TCP/IP protocol suite, which supports popular Internet applications such as web browsing, ftp, and e-mail. The Internet protocol as it is does not support mobility, and therefore cannot be successfully used in a wireless environment. On the other hand, due to the widespread penetration of the World Wide Web, it is impossible to change the protocols in use in the whole Internet, and considerable research has gone into studying how to devise schemes which can support mobility while being compatible with the protocols currently implemented throughout the world. An example is Mobile IP, a version of the Internet protocol which supports host mobility (but not handoff) through the provision of a care-of address capability used by the "home network" to forward traffic for one of its hosts which may temporarily reside in a different ("foreign") network (Perkins, 1998).

Another important problem to be solved before popular computer applications find their way in the wireless domain concerns the suitability of TCP as a reliable transport protocol over connections comprising wireless segments. It has been reported in several papers (e.g., see Balakrishnan *et al.*, 1997b and the references therein) that current releases of TCP do not work properly on wireless channels because they misinterpret packet losses due to channel impairments and assume that congestion occurred. The action triggered by the protocol rules in this case simply addresses the problem in the wrong way, and catastrophic throughput losses have been observed. Possible solutions to this problem include changing the rules of TCP (with serious compatibility problems), terminating the TCP connection at the BS, (which violates the semantics of TCP which is an end-to-end protocol), and introducing some link control mechanism to enhance the quality of the radio channel (see Balakrishnan *et al.*, 1997b for a comparative study of various schemes). In particular, the last solution is attractive since it does not involve any change to current TCP implementations, and relies on ARQ techniques which are able to recover erroneous packets before a timeout expires and avoid congestion control protocol actions (Chaskar *et al.*, 1996). Other issues related to the use of TCP on wireless channels are discussed in Balakrishnan *et al.*, 1997a. The performance of TCP with an underlying radio link protocol on DS-CDMA wireless links is presented in Chockalingam and Bao, 2000. A unified approach to analyzing TCP throughput performance on wireless fading links with memory is presented in (Zorzi *et al.*, 2000).

E. Energy and Power Management

Battery life has always been a concern in the wireless communications industry, since improved battery performance translates in smaller and lighter devices. Unfortunately,

the rate at which battery performance improves (in terms of available energy per unit size or weight) is fairly slow, despite the great interest generated by the booming wireless business. Other means to reduce power consumption in wireless devices have therefore been receiving interest in the research community. Besides the traditional search for better circuits and power amplifiers, which directly improve the power consumption efficiency of the whole device, some recent research results have taken a broader view of the problem. In particular, it has been recognized that energy conservation is a task which can be performed at multiple levels, and energy-efficient protocol design criteria can be identified and applied. This concept has resulted in studies on energy-efficient error control schemes (where bad channel conditions cause the transmitter to enter a sleep mode), multiple-access protocols (by minimizing collisions and the need for retransmission), software techniques (such as various low-power modes at the operating system level), and signal processing algorithms (via more careful software circuit design). Another traditional topic for wireless communications, namely power control, has been recently revisited in a network resource allocation perspective (as opposed to simple attenuation compensation), opening up new optimization problems and networking issues. Advances on these topics can be found in (IEEE, 1995a, 1998; Bambos and Rulnick, 1997; Sivalingam *et al.*, 1997; Zorzi and Rao, 1997; and Chockalingam and Zorzi, 1998).

F. Toward Multimedia Radios

In the above description, adaptive techniques have been mentioned several times as a means to face the erratic behavior of the wireless channel. Also, adaptation of applications to changing network conditions may be a feasible means to overcome such difficulty without compromising the network efficiency.

Another good reason for adaptivity stems from the fact that applications may have widely different needs, and choices which are good (or even optimal) for one type of service may be suboptimal (or even catastrophic) for others. The current wireless infrastructure has been designed based on the assumption of (uniform) voice traffic, and many communications functions have been optimized under these conditions. It may be expected that with the emergence of data services this kind of network infrastructure will not be adequate. For example, packet switching needs to replace circuit switching. Error control, performed by FEC and interleaving in current voice systems, may be done more efficiently by ARQ schemes for many data applications. There is evidence that some services may tolerate correlated errors better than dispersed errors (while the opposite is true for current voice communication systems) (Zorzi and Rao, 1997c). Also, it is likely that there exists no solution which is optimal for everything, but rather various solutions which best address the needs of specific classes of applications.

In current implementations, being able to reach down to the radio level and to access and reprogram the hardware is very hard, and in many cases impossible. However, for a flexible device which can be used to support a variety of applications, ranging from web browsing to plain telephony and from e-mail to full motion video transmission, being able to select the transmission scheme which best adapts to the specific application needs may have a great value. Besides giving the possibility of selecting at any time the best possible scheme (e.g., by being able to turn off some blocks, or to select the modulation/error control scheme by simple software instructions), such a "software radio" could be upgraded easily via software downloading. The military started thinking about this a few years ago, due to the need to achieve interoperability across different frequency bands, channel modulations, and data formats. Commercial interest has been growing as well in recent years (IEEE, 1995b).

Even protocol stacks could be programmed and their functionalities negotiated, so that the user (or the application, without need for direct human intervention) can select the protocol specifications which best suit the needs of the service being sought. The negotiation process would then extend to protocol specifications, beyond a simple set of connection parameters. The paradigm of intelligent software agents and Java, where code segments and modules can be downloaded on the fly, may apply in the context of communications systems as well.

This vision promises full liberation of radio systems from the slavery of "hard-wired" features, including data rates, channel coding, and frequency bands, and is based on the use of multiband antennas, wideband A/D and D/A conversions, and digital signal-processing capabilities through general-purpose programmable processors. This new radio architectural concept (which can reach up to higher layers, as briefly described before) will lead to more efficient systems, where no hardware designer makes choices *a priori* for the user or the application. Truly universal services and devices will then be realized.

VI. FURTHER READING

The purpose of this article was essentially to give a broad overview of the issues involved in the provision of telecommunications services in wireless and mobile networks, by describing the key technical challenges, current system designs, and future trends. Because of the great

variety of issues and the large amount of research which has been done on this topic (fueled by the tremendous interest in wireless communications), it is clearly impossible to give any level of detail in such a limited space. The interested reader will find many in-depth books and technical articles which address many specific aspects of this extremely rich field. Books on the fundamentals of mobile communications and of cellular systems include Garg and Wilkes, 1995b; Goodman, 1997; Jakes, Jr., 1974; Lee, 1993; Pahlavan and Levesque, 1995; and Rappaport, 1996. Tutorial articles covering various aspects of this topic and reporting extensive literature citations include Acampora, 1996; Chen, 1994; Cox, 1995; Gilhousen *et al.*, 1991; Hashemi, 1993; Li and Qiu, 1995; Katzela and Naghshineh, 1996; LaMaire *et al.*, 1996; Pahlavan and Levesque, 1994, Pickholtz *et al.*, 1991; and Schwartz, 1995.

Articles dealing with various aspects of wireless communications systems can be found in the technical literature. Journals such as the *IEEE Transactions on Communications*, the *IEEE Journal on Selected Areas in Communications* (which publishes single-topic special issues), and the *IEEE Communications Magazine* (intended for a broader audience and more of a tutorial nature) deal with all aspects of communication theory and networking. Periodicals entirely devoted to research on wireless communications and networks include the *IEEE Transactions on Vehicular Technology*, the *IEEE Personal Communications* Magazine, the *ACM/Baltzer Journal of Wireless Networks*, and several other new journals started in the past few years to collect the many technical contributions in the field. Many conferences on wireless systems are also held every year.

APPENDIX: LIST OF ACRONYMS

ACCH	Associated Control Channels
ADPCM	Adaptive Differential Pulse Code Modulation
AMPS	Advanced Mobile Phone Service
AJ	Anti-Jamming
ARDIS	Advanced Radio Data Information System
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
BPSK	Binary Phase Shift Keying
BS	Base Station
CCITT	International Telegraph and Telephone Consultative Committee
CDMA	Code Division Multiple Access
CDPD	Cellular Digital Packet Data
CPCH	Common Packet CHannel
DCH	Dedicated Channel
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
DCS1800	Digital Communications System—1800
DECT	Digital European (Enhanced) Cordless Telecommunication
DIFS	Distributed-Coordination-Function IFS
DQPSK	Differential Quadrature Phase Shift Keying
DS-CDMA	Direct Sequence CDMA
DSCH	Downlink Shared Channel
DSMA/CD	Digital Sense Multiple Access with Collision Detection
EDGE	Enhanced data Rates for Global Evolution
FACH	Forward Access Channel
FCC	Federal Communications Commission
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
FEC	Forward Error Correction
FM	Frequency Modulation
FPLMTS	Future Public Land Mobile Telecommunications Systems
F-SCH	Forward Supplemental CHannel
GEO	Geostationary Orbit
GGSN	Gateway GPRS Support Node
GMSK	Gaussian Minimum Shift Keying
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communication
HDML	Handheld Device Markup Language
HDTP	Handheld Device Transfer Protocol
HIPERLAN	High Performance Radio LAN
HLR	Home Location Register
HSCSD	High Speed Circuit Switched Data
HTML	HyperText Markup Language
IFS	Interframe Space
ISDN	Integrated Services Digital Network
LAN	Local Area Network
LEO	Low Earth Orbit
LLC	Logical Link Control
LPI	Low Probability of Intercept

MAC	Media Access Control
MAHO	Mobile Assisted HandOff
MAI	Multiple Access Interference
MEO	Medium Earth Orbit
MF	Matched Filter
MSC	Mobile Switching Center
MSK	Minimum Shift Keying
OFDM	Orthogonal Frequency Division Multiplexing
PACS	Personal Access Communication System
PAN	Personal Area Networks
PBX	Private Branch Exchange
PCS	Personal Communications Services
PDC	Pacific Digital Cellular
PIFS	Point-Coordination-Function IFS
PRMA	Packet Reservation Multiple Access
PSTN	Public Switched Telephone Network
QPSK	Quadrature Phase Shift Keying
QoS	Quality of Service
RLC	Radio Link Control
RLP	Radio Link Protocol
RPE	Regular Pulse Excited
R-SCH	Reverse Supplememtal CHannel
SACCH	Slow Associated Control Channels
SGSN	Serving GPRS Support Node
SIFS	Short Interframe Space
TACS	Total Access Communications System
TCM	Trellis Coded Modulation
TCP	Transport Control Protocol
TDD	Time Division Duplexing
TDMA	Time Division Multiple Access
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications Services
VLR	Visitor Location Register
VSELP	Vector Sum Excited Linear Predictor
WAE	Wireless Application Environment
WAP	Wireless Application Protocol
WCDMA	Wideband CDMA
WDP	Wireless Datagram Protocol
WLAN	Wireless Local Area Network
WML	Wireless Markup Language
WSP	Wireless Session Protocol
WTLS	Wireless Transport Layer Security
WTP	Wireless Transaction Protocol

SEE ALSO THE FOLLOWING ARTICLES

BASEBAND DIGITAL LOOPS • COMPUTER NETWORKS • DESIGNS AND ERROR-CORRECTING CODES • NETWORKS FOR DATA COMMUNICATION • RADIO PROPAGATION • TELECOMMUNICATIONS • WWW (WORLD-WIDE WEB)

BIBLIOGRAPHY

Proc. IEEE (1995). special issue on "Low-power electronics," April.

IEEE Commun. Mag. (1995). special issue on "Software radios."

IEEE Personal Commun. Mag. (1998). special issue on "Energy management in personal communications/mobile computing."

Acampora, A. S. (1994). "An Introduction to Broadband Networks: LANs, MANs, ATM, B-ISDN, and Optical Networks for Integrated Multimedia Telecommunications," Plenum Press, New York.

Acampora, A. S., and Naghshineh, M. (1994). An architecture and methodology for mobile-executed handoff in cellular ATM networks. *IEEE J. Selected Areas Commun.* **12,** 1365–75.

Acampora, A. S. (1996). "Wireless ATM: a perspective on issues and prospects," *IEEE Personal Commu.* **3,** 8–17.

Acampora, A. S., Krishnamurthy, S., and Zorzi, M. (1997). "Media Access Protocols for use with Smart Array Antennas to Enable Wireless Multimedia Applications," *In* Proc. 9th Tyrrhenian Workshop on Digital Communications, September.

Ayanoglu, E., Eng, K. Y., and Karol, M. J. (1996). "Wireless ATM: limits, challenges, and proposals," *IEEE Personal Commun.* **3,** 18–34.

Balakrishnan, H., *et al.* (1997). TCP improvements for heterogeneous networks: the Daedalus approach. *In* Proc. 35th Annual Allerton Conference.

Balakrishnan, H., Padmanabhan, V. N., Seshan, S., and Katz, R. H. (1997). A comparison of mechanisms for improving TCP performance over wireless links. "ACM/IEEE Trans. on Networking."

Bambos, N., and Rulnick, J. M.(1997). Mobile power management for wireless communication networks. *Wireless Networks* **3,** 3–14.

Berrou, C., Glavieux, A., and Thitimajshima, P. (1993). Near Shannon limit error correcting coding and decoding: turbo codes. "Proc. IEEE ICC'93," pp. 1064–1070. Also in *IEEE Trans. Commun.* **44,** pp. 1261–71.

Bettstetter, C., Vogel, H. J., and Eberspacher, J. (1999). GSM Phase 2+ general packet radio service GPRS: architecture, protocols, and air interface. *IEEE Commun. Surveys*, pp. 2–14, Vol. 2, No. 3, 3rd Quarter.

Bluetooth Specification, http://www.bluetooth.com/.

Bianchi, G., Borgonovo, F., Fratta, L., Musumeci, L., and Zorzi, M. (1997). C-PRMA: a centralized packet reservation multiple access for local wireless communications. *IEEE Trans. Veh. Tech.* **46,** 422–436.

Biglieri, E., *et al.* (1991). "Introduction to Trellis-Coded Modulation, with Applications," Macmillan, New York.

Bingham, J. (1990). Multicarrier modulation for data transmission: an idea whose time has come. *IEEE Commun. Mag.* **28,** 5–14.

Borgonovo, F., Fratta, L., Zorzi, M., and Acampora, A. S. (1996). Capture division packet access—A new cellular access architecture for future PCNs. *IEEE Commun. Mag.* **34,** No. 9, 154–162.

Chaskar, H., Lakshman,T. V., and Madhow, U. (1996)."On the design of interfaces for TCP/IP over wireless," *In Proc. MILCOM'96*.

Chen, K. C. (1994). "Medium access control of wireless LANs for mobile computing," *IEEE Network* **8,** 50–63.

Chockalingam, A., and Zorzi, M. (1998). Energy efficiency of media access protocols for mobile data networks. *IEEE Trans. Commun.* **46,** No. 11, 1418–1421.

Chockalingam, A., and Bao, G. (2000). Performance of TCP/RLP protocol stack on correlated fading DS-CDMA wireless links. *IEEE Trans. Veh. Technol.* **49,** No. 1, 28–33.

Cox, D. (1995). Wireless personal communications: what is it?. *IEEE Personal Commun.* **2,** 20–35.

Falconer, "A system architecture for broadband millimeter-wave access to an ATM LAN," *IEEE Personal Communications* vol. 3, pp. 36–41, Aug. 1996.

Foschini, G. J., and Gans, M. J. (1998). On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Commun.* **6,** 311–35.

Fukasawa, A., *et al.* (1996). Wideband CDMA system for personal radio communications. *IEEE Commun. Mag.* **34,** No.10, 116–123.

Furuskar, A., Mazur, S., Muller, F., and Olofsson, H. (1999). EDGE: Enhanced data rates for GSM and TDMA/136 evolution. *IEEE Personal Commun.*, 56–66.

Garg, V. K. (2000). "IS-95 CDMA and cdma2000: Cellular/PCS Systems Implementation," Prentice-Hall, Englewood Cliffs, New Jersey.

Garg, V. K., and Wilkes, J. E. (1996). " Wireless and Personal Communications Systems," Prentice Hall, Upper Saddle River, New Jersy.

Giannakis, G. B., Hua, Y., Stoica, P., and Tong, L. (2000). "Signal Processing Advances in Wireless & Mobile Communications: Trends in Single- and Multi-User Systems, Vol. 2," Prentice-Hall, New Jersey.

Gilhousen, K. S., *et al.* (1991). On the capacity of a cellular CDMA system. *IEEE Trans. Veh. Tech.* **VT-40,** 303–312.

Goodman, D. J., *et al.* (1989). Packet reservation multiple access for local wireless communications. *IEEE Trans. Commun.* **37,** 885–890.

Goodman, D. J. (1997). "Wireless Personal Communications Systems," Addison-Wesley, Reading, MA.

Grandhi, S. A., Yates, R. D., and Goodman, D. J. (1997). Resource allocation for cellular radio systems. *IEEE Trans. Veh. Tech.* **VT-46,** 581–588.

Hashemi, H. (1993). The indoor radio propagation channel. *Proc. IEEE* **81,** 943–68.

Holma, H., and Toskala, A. (2000). "WCDMA for UMTS: Radio Access for Third Generation Mobile Communications," John-Wiley, New York.

Jakes, Jr., W. C. (1974). "Microwave Mobile Communications," John Wiley & Sons, New York.

Karol, M. J., Zhao L., and Eng, K. Y. (1995). "Distributed-queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks." *In Proc. ICC'95*, pp. 1224–31.

Katzela, I., and Naghshineh, M. (1996). Channel assignment schemes for cellular mobile telecommunication systems: a comprehensive survey. *IEEE Personal Commun.* **3,** 10–31.

Khansari, M., *et al.* (1996). Low bit-rate video transmission over fading channels for wireless microcellular systems. *IEEE Trans. Circuits Systems* Video Techno.* **6,** 1–11.

LaMaire, R. O., *et al.* (1996). Wireless LANs and mobile networking: standards and future directions. *IEEE Commun. Mag.* **34,** 86–94.

Lee, W. C. Y. (1993). "Mobile Communications Design Fundamentals," 2nd ed., Wiley, New York.

Li, V. O. K., and Qiu, X. (1995). Personal communication systems (PCS). *Proc. IEEE* **83,** 1210–43.

Lin, S., and Costello, D. J. (1983). "Error Control Coding: Fundamentals and Applications," Prentice-Hall, Englewood Cliffs, New Jersey.

Mitrou, N. M., Orinos, T. D., and Protonotarios, E. N. (1990). A reservation multiple access protocol for microcellular mobile-communication system. *IEEE Trans. Veh. Technol.* **VT-39,** 340–351.

Naghshineh, M., and Acampora, A. S. (1996). QoS provisioning in micro-cellular networks supporting multiple classes of traffic. *Wireless Networks* **2,** 195–203.

Naghshineh, M., and Willebeek-LeMair, M. (1997). End to end QoS provisioning multimedia wireless/mobile networks using an adaptive framework *IEEE Commun. Mag.* special issue on "Introduction to mobile and wireless ATM."

Onvural, R. (1995). "Asynchronous Transfer Mode Networks: Performance Issues," Artech House, Boston.

Pahlavan, K., and Levesque, A. H. (1995). "Wireless Information Networks," Wiley, New York.

Pahlavan, K., and Levesque, A. H. (1994). Wireless data communcations. *Proc. IEEE* **82,** 1398–1430.

Perkins, C. (1998). "Mobile IP: Design Principles and Practices," Addison-Wesley, Reading, MA.

Pickholtz, R. L., Milstein, L. B., and Schilling, D. L., (1991). Spread spectrum for mobile communications. *IEEE Trans. Veh. Technol.* **40,** No. 2, 313–322.

Rappaport, T. S. (1996). "Wireless Communications: Principles and Practice," Prentice-Hall, Upper Saddle River, NJ.

Schilling, D. L., *et al.* (1991). Broadband CDMA for personal communications systems. *IEEE Commun. Mag.* **29,** 86–93.

Schwartz, M. (1995). Network management and control issues in multimedia wireless networks. *IEEE Personal Commun.* **2,** 8–16.

Schwartz, M. (1996). "Broadband Integrated Networks," Prentice-Hall, Upper Saddle River, NJ.

Sivalingam, K. M., Srivastava, M. B., and Agrawal, P. (1997). Low power link and access protocols for wireless multimedia networks. *In Proc. IEEE VTC'97*, pp. 1331–35.

Tarokh, V., Seshadri, N., and Calderbank, A. R. (1998). Space-time codes for high data rate wireless communication: performance criterion and code construction. *IEEE Trans. Information Theory* **44,** 744–65.

TIA/EIA/IS-95 (1993). Mobile Station-Base Station Compatibility Standard for Dual-Mode Wideband Spectrum Cellular System.

TIA/EIA/IS-99 (1999). Data Services Option Standard for Wideband Spread Spectrum Cellular System.

Veeraraghavan, M., Karol, M., and Eng, K. (1997). Mobility and connection management in a wireless ATM LAN. *IEEE Selected Areas Commun.* **15,** 50–68.

Verdu, S. (1998). " Multiuser Detectio," Cambridge Univ. Press, U.K.

Viterbi, A. J. (1995). "CDMA: Principles of Spread Spectrum Communication," Addison-Wesley, Reading, MA.

Wu, G., Mukumoto, K., and Fukuda, A. (1994). Performance evaluation of reserved idle signal multiple access scheme for local wireless communications. *IEEE Trans. Veh. Techno.* **VT-43,** 653–658.

Wu, T. H., and Chang, L. (1995). Architectures for PCS mobility management on ATM transport networks. *1995 Fourth IEEE Int. Conf. Universal Personal Commun.*

Xie, Z., Short, R. T., and Rushforth, C. K. (1990). A family of suboptimum detectors for coherent multiuser communications. *IEEE Jl. Sel. Areas Commun.* **8,** 683–690.

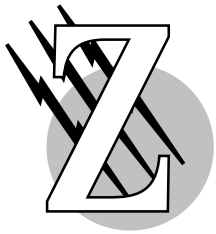
Zorzi, M., and Pupolin, S. (1994). Slotted ALOHA for high capacity voice cellular communications. *IEEE Trans. Veh. Technol.* **43,** 1011–1021.

Zorzi, M., and Rao, R. R. (1997). Error control and energy consumption in communications for nomadic computing. *IEEE Trans. Computers* (special issue on Mobile Computing) **46,** 279–289.

Zorzi, M., and Rao, R. R. (1997). "ARQ error control for delay-constrained communications on short-range burst-error channels," *In Proc. VTC'97*, pp. 1528–32.

Zorzi, M., and Rao, R. R. (1997). The role of error correlations in the design of protocols for packet switched services. *In Proc. 35th Annual Allerton Conf.*

Zorzi, M., Chockalingam, A., and Rao, R. R. (2000). .Throughput Analysis of TCP on Channels with Memory. *IEEE Jl. Sel. Areas Commun. (JSAC)—Wireless Commun. Series* Vol. 18, No. 7, pp. 1289–1300.



z-Transform

M. Edwin Sawan
Wichita State University

I. Definition
II. Inverse *z*-Transform
III. Transfer Function
IV. Relation to Laplace Transform
V. Limitations of the *z*-Transform
VI. Delta Transform

GLOSSARY

Causal signal Time function that is equal to zero for negative values of time.
Complex variable Complex quantity where the real and imaginary parts are represented with independent variables.
Convolution Mathematical formula that describes the time domain relationship between the input and output of a linear system. Here, it is a summation that involves the system's impulse response.
Difference equation Equation relating advanced and/or delayed versions of time functions. It is used to describe the interaction of components of discrete-time dynamic systems.
Dynamic system System with energy-storing components. The order of the system is determined by the number of such components.
Mapping Transformation according to a specific formula from one set to another.
Pole Value of a variable that causes the denominator of a rational function to equal zero.
Sampling period Time interval between two measurements of a time function.
Simulation diagram Sketch of a system that illustrates the interactions among its components, where dynamic elements are represented by simple integrators or unit delays.
Transfer function Ratio between the z-transforms of the input and output of a linear system assuming zero initial conditions.

THE *z*-TRANSFORM is a mathematical tool used for the analysis and design of discrete systems. Its basic role is the transformation of linear difference equations into linear algebraic equations by means of mapping discrete-time signals from the time domain into the so-called z-domain, where an alternate representation is given in terms of a complex variable z. Hence, a dynamic system can be represented in the z-domain by a transfer function that is equal to the ratio between the input and output. Because of the algebraic nature of this representation, it enhances the modeling, analysis, and design of discrete-time systems. Application of the z-transform technique is supported by comprehensive tables of transform pairs for common signals, properties, and theorems.

I. DEFINITION

The z-transform of a causal signal $x(k)$, where k takes zero or positive integer values, is defined by the following equation:

$$X(z) = \sum_{k=0}^{\infty} x(k)z^{-k}.$$

For a noncausal signal $x(k)$ where k takes integer values ($k = \ldots, -2, -1, 0, 1, 2, \ldots$), the two-sided z-transform is defined by the following equation:

$$X(z) = \sum_{k=-\infty}^{\infty} x(k)z^{-k}.$$

Being an infinite series, the z-transform may or may not converge, depending on the magnitude of z. For any signal $x(k)$, there exists a region in the z-domain for which the z-transform converges. The following are some basic properties based on given z-transform pairs, $Z\{f(k)\} = F(z)$ and $Z\{g(k)\} = G(z)$:

1. *Linearity*.

$$Z\{af(k) + bg(k)\} = aF(z) + bG(z)$$

where a and b are any constant multipliers.

2. *Translation*.

$$Z\{f(k+N)\} = z^N F(z) - [z^N f(0) + z^{N-1} f(1) + \cdots + zf(N-1)]$$

where N is an integer number.

3. *Convolution*. The convolution sum of two causal signals f and g is defined as follows:

$$f(k) \cdot g(k) = \sum_{j=0}^{k} f(j)g(k-j)$$

or, equivalently,

$$f(k) \cdot g(k) = \sum_{j=0}^{k} f(k-j)g(j).$$

The z-transform of the convolution sum is given by

$$Z\{f(k) \cdot g(k)\} = F(z)G(z)$$

4. *Initial value*.

$$f(0) = \lim_{z\to\infty} F(z)$$

5. *Final value*.

$$f(\infty) = \lim_{z\to 1}(1 - z^{-1})F(z)$$

provided that the poles of $(1 - z^{-1})F(z)$ are strictly inside the unit circle of the complex z-plane.

6. *Multiplication by an exponential factor* α^k.

$$Z\{\alpha^k f(k)\} = F(z/\alpha)$$

7. *Summation*.

$$Z\left\{\sum_{i=0}^{k} f(i)\right\} = \frac{z}{z-1} F(z)$$

8. *Frequency differentiation*.

$$Z\{kf(k)\} = -z\frac{d}{dz} F(z)$$

9. *Frequency integration*.

$$Z\left\{\frac{1}{k} f(k)\right\} = \int_z^{\infty} \frac{1}{w} F(w)\, dw$$

II. INVERSE z-TRANSFORM

There are several methods to compute the inverse z-transform.

A. Inversion Integral Formula

The inversion integral for the z-transform $X(z)$ is given by the following equation:

$$x(k) = \left(\frac{1}{2\pi j}\right) \oint_C X(z)z^{k-1}\, dz,$$

where C is a circle with its center at the origin of the z-plane such that all poles of $X(z)z^{k-1}$ are inside it.

B. Partial Fraction Expansion

A function $F(z)$ is expanded in partial fractions. Then, tables of known z-transform pairs are used to determine the inverse z-transform. Typically, the partial fraction expansion is performed on $(1/z)F(z)$. This will result in the terms of $F(z)$ being of the same form as those in the tables since a factor of z is generally required in the numerator at the z-transform table. The following are examples of commonly used transform pairs:

$$\begin{aligned}
f(k) &= 1 \quad \leftrightarrow F(z) = z/(z-1)\\
f(k) &= k \quad \leftrightarrow F(z) = z/(z-1)^2\\
f(k) &= k^2 \quad \leftrightarrow F(z) = z(z+1)/(z-1)^3\\
f(k) &= a^k \quad \leftrightarrow F(z) = z/(z-a)\\
f(k) &= ka^k \leftrightarrow F(z) = az/(z-a)^2.
\end{aligned}$$

C. Power Series Expansion

The denominator of $F(z)$ is divided into the numerator. Using long division, we obtain a power series of the following form:

$$F(z) = f_0 + f_1 z^{-1} + f_2 z^{-2} + f_3 z^{-3} + \cdots$$

From the definition of the z-transform, it is clearly seen that the values of $f(k)$ are simply the coefficients in the power series.

III. TRANSFER FUNCTION

One of the basic properties of the z-transform is that convolution is transformed into multiplication. This gives rise to the transfer function definition as the ratio between the output and the input of a dynamic system under zero initial conditions. Transfer function is a useful tool of modeling, analysis, and design of dynamic systems. Its poles, also known as the characteristic roots, determine the system stability. A system is stable if all its poles lie inside the unit circle centered at the origin of the z-plane. Transfer functions are also used to develop block simulation diagrams that are essential for the analysis and design of dynamic systems.

There are several methods to test the stability of discrete-time systems without actually computing the poles of the transfer function. A frequently used method is based on the Routh–Hurwitz criterion for continuous systems, where stability is indicated by transfer function poles lying on the left half of the complex plane. The method starts with applying the bilinear transformation

$$z = (w + 1)/(w - 1)$$

to the discrete transfer function. This transformation maps the transfer function from the z-plane to another complex plane, the w-plane. It is clearly seen that the inside of the unit circle in the z-plane is mapped into the left half of the w-plane. Hence, the Routh–Hurwitz stability criterion can be used to determine the stability of the transfer function represented in the w-plane.

IV. RELATION TO LAPLACE TRANSFORM

The z-transform is regarded as the discretetime equivalence of the Laplace transform. The relation between the two transforms is represented by the following mapping:

$$z = e^{s\Delta},$$

where s is the Laplace transform variable and Δ is the sampling period of the signal. The left half of the s-plane is mapped onto a circle with radius = 1, centered at the origin of the z-plane.

The relationship between the z-transform and the Laplace transform can be illustrated by considering a continuous time function $x(t)$ which is sampled by an ideal sampler. The resulting sampled-data signal is given by the following:

$$x^*(t) = \sum_{k=0}^{\infty} x(k\Delta)\delta(k\Delta),$$

where $\delta(\cdot)$ is the impulse function.

Then the Laplace transform of $x^*(t)$ is computed as follows:

$$X^*(s) = \mathcal{L}\{x^*(t)\} = \sum_{k=0}^{\infty} x(k\Delta)e^{-k\Delta s}.$$

Finally, $e^{\Delta s}$ is replaced with z in $X^*(s)$, resulting in the z-transform as follows:

$$X(z) = \sum_{k=0}^{\infty} x(k\Delta)z^{-k}.$$

V. LIMITATIONS OF THE *z*-TRANSFORM

The definition and properties of the z-transform are based on certain assumptions. Hence, care must be taken in the application and the interpretation of the results of the z-transform methods. Special attention must be given to the cases where the z-transform is applied to sampled signals. Such application is valid only if the sampling period is small when compared with the significant time constant of the system. Furthermore, the application of the z-transform will result in losing information about the signal between the sampling instants since the z-transform describes the time function only at the sampling instants.

VI. DELTA TRANSFORM

In order to establish a unified transform theory, the z-transform is scaled by the sampling period Δ and an alternate transform variable γ, given by

$$\gamma = \frac{z - 1}{\Delta}$$

is used. This gives rise to the definition of the so-called δ-transform as follows:

$$F_\Delta(\gamma) = \Delta \sum_{k=0}^{\infty} f(k)[1 + \Delta\gamma]^{-k}.$$

A table of transforms can readily be obtained for delta transforms by simply noting that

$$F_\Delta(\gamma) = \Delta F(z)|_{z=\Delta\gamma+1}.$$

Properties and theorems of the delta transform closely follow those of Laplace transform for continuous systems, thus providing an insightful way to describe the sampled-data case.

SEE ALSO THE FOLLOWING ARTICLES

APPROXIMATIONS AND EXPANSIONS • DISCRETE SYSTEMS MODELING • LINEAR SYSTEMS OF EQUATIONS • SIMULATION AND MODELING • SYSTEM THEORY

BIBLIOGRAPHY

Chen, C.-T. (1993). "Analog and Digital Control System Design, Transfer-Function, State-Space, and Algebraic Methods," Oxford University Press, Oxford, UK.

Franklin, P. (1998). "Digital Control of Dynamic Systems," 3rd ed., Addison-Wesley.

Jacquot, R. G. (1994). "Modern Digital Control Systems," 2nd ed., Marcel Dekker, New York.

Nekoogar, F., and Moriarty, G. (1998). "Digital Control Using Digital Signal Processing," Prentice Hall, New York.

Phillips, C. L., and Nagle, T. H. (1994). "Digital Control System Analysis and Design," Prentice Hall.

Vaccaro, R. (1995). "Digital Control," McGraw–Hill, New York.