

Acoustic Chaos

Werner Lauterborn
Universität Göttingen

I. The Problem of Acoustic Cavitation Noise
II. The Period-Doubling Noise Sequence
III. A Fractal Noise Attractor
IV. Lyapunov Analysis
V. Period-Doubling Bubble Oscillations
VI. Theory of Driven Bubbles
VII. Other Systems
VIII. Philosophical Implications

GLOSSARY

Bifurcation Qualitative change in the behavior of a system, when a parameter (temperature, pressure, etc.) is altered (e.g., period-doubling bifurcation); related to phase change in thermodynamics.

Cavitation Rupture of liquids when subject to tension either in flow fields (hydraulic cavitation) or by an acoustic wave (acoustic cavitation).

Chaos Behavior (motion) with all signs of statistics despite an underlying deterministic law (often, deterministic chaos).

Fractal Object (set of points) that does not have a smooth structure with an integer dimension (e.g., three dimensional). Instead, a fractal (noninteger) dimension must be ascribed to them.

Period doubling Special way of obtaining chaotic (irregular) motion; the period of a periodic motion doubles repeatedly until in the limit of infinite doubling aperiodic motion is obtained.

Phase space Space spanned by the dependent variables of a dynamic system. A point in phase space characterizes a specific state of the system.

Strange attractor In dissipative systems, the motion tends to certain limits forms (attractors). When the motion comes to rest, this attractor is called a fixed point. Chaotic motions run on a strange attractor, which has involved properties (e.g., a fractal dimension).

THE PAST FEW years have seen a remarkable development in physics, which may be described as the upsurge of "chaos." Chaos is a term scientists have adapted from common language to describe the motion or behavior of a system (physical or biological) that, although governed by an underlying deterministic law, is irregular and, in the long term, unpredictable.

Chaotic motion seems to appear in any sufficiently complex dynamical system. Acoustics, that part of physics that descibes the vibration of usually larger ensembles of molecules in gases, liquids, and solids, makes no exception. As a main necessary ingredient of chaotic

dynamics is nonlinearity, acoustic chaos is closely related to nonlinear oscillations and waves in gases, liquids, and solids. It is the science of never-repeating sound waves. This property it shares with noise, a term having its origin in acoustics and formerly attributed to every sound signal with a broadband Fourier spectrum. But Fourier analysis is especially adapted to linear oscillatory systems. The standard interpretation of the lines in a Fourier spectrum is that each line corresponds to a (linear) mode of vibration and a degree of freedom of the system. However, as examples from chaos physics show, a broadband spectrum can already be obtained with just three (nonlinear) degrees of freedom (that is, three dependent variables). Chaos physics thus develops a totally new view of the noise problem. It is a deterministic view, but it is still an open question how far the new approach will reach in explaining still unsolved noise problems (e.g., the $1/f$-noise spectrum encountered so often). The detailed relationship between chaos and noise is still an area of active research. An example, where the properties of acoustic noise could be related to chaotic dynamics, is given below for the case of acoustic cavitation noise.

Acoustic chaos appears in an experiment when a liquid is irradiated with sound of high intensity. The liquid then ruptures to form bubbles or cavities (almost empty bubbles). The phenomenon is known as acoustic cavitation and is accompanied by intense noise emission—the acoustic cavitation noise. It has its origin in the bubbles set into oscillation in the sound field. Bubbles are nonlinear oscillators, and it can be shown both experimentally and theoretically that they exhibit chaotic oscillations after a series of period doublings. The acoustic emission from these bubbles is then a chaotic sound wave (i.e., irregular and never repeats). This is acoustic chaos.

I. THE PROBLEM OF ACOUSTIC CAVITATION NOISE

The projection of high-intensity sound into liquids has been investigated since the application of sound to locate objects under water became used. It was soon noticed that at too high an intensity the liquid may rupture, giving rise to acoustic cavitation. This phenomenon is accompanied by broadband noise emission, which is detrimental to the useful operation of, for instance, a sonar device.

The noise emission presents an interesting physical problem that may be formulated in the following way. A sound wave of a single frequency (a pure tone) is transformed into a broadband sound spectrum, consisting of an (almost) infinite number of neighboring frequencies. What is the physical mechanism that causes this transformation? The question may even be shifted in its emphasis to ask what physical mechanisms are known to convert a single frequency to a broadband spectrum? This could not be answered before chaos theory was developed. However, although chaos theory is now well established, a physical (intuitive) understanding is still lacking.

II. THE PERIOD-DOUBLING NOISE SEQUENCE

To investigate the sound emission from acoustic cavitation the experimental arrangement as depicted in Fig. 1 is used. To irradiate the liquid (water) a piezoceramic cylinder (PZT-4) of 76-mm length, 76-mm inner diameter, and 5-mm wall thickness is used. When driven at its main resonance, 23.56 kHz, a high-intensity acoustic field is generated in the interior and cavitation is easily achieved. The noise is picked up by a broadband hydrophone and digitized at rates up to 60 MHz after suitable lowpass filtering (for correct analog-to-digital conversion for later processing) and strong filtering of the driving frequency, which would otherwise dominate the noise output. The experiment is fully computer controlled. The amplitude of the driving sound field can be made an arbitrary function of time via a programmable synthesizer. In most cases, linear ramp functions are applied to study the buildup of noise when the driving pressure amplitude in the liquid is increased.

From the data stored in the memory of the transient recorder, power spectra are calculated via the fast-Fourier-transform algorithm from usually 4096 samples out of the 128×1024 samples stored. This yields about 1000 short-time spectra when the 4096 samples are shifted by 128 samples from one spectrum to the next.

Figure 2 shows four power spectra from one such experiment. Each diagram gives the excitation level at

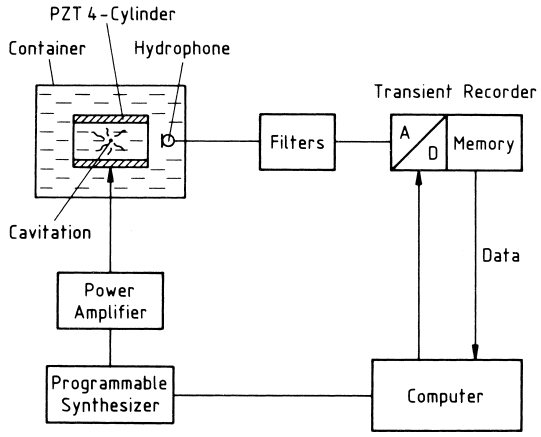


FIGURE 1 Experimental arrangement for measurements on acoustic cavitation noise (chaotic sound).

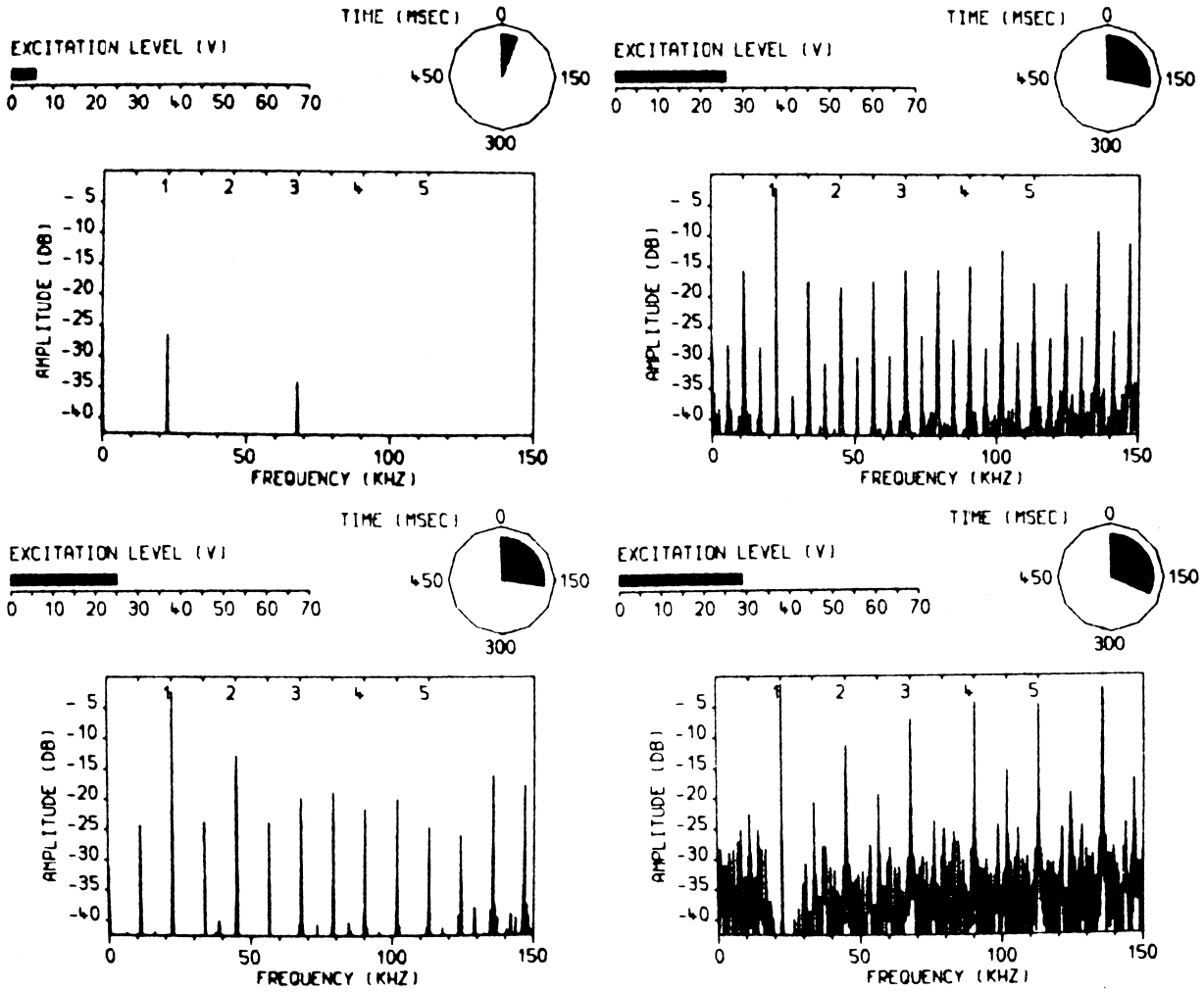


FIGURE 2 Power spectra of acoustic cavitation noise at different excitation levels (related to the pressure amplitudes of the driving sound wave). (From Lauterborn, W. (1986). *Phys. Today* **39,** S-4.)

the transducer in volts, the time since the experiment (irradiating the liquid with a linear ramp of increasing excitation) has started in milliseconds, and the power spectrum at this time. At the beginning of the experiment, at low sound intensity, only the driving frequency f_0 shows up. In the upper left diagram of Fig. 2 the third harmonic, $3f_0$, is present. When comparing both lines it should be remembered that the driving frequency is strongly damped by filtering. In the lower left-hand diagram, many more lines are present. Of special interest is the spectral line at $\frac{1}{2}f_0$ (and their harmonics). A well-known feature of nonlinear systems is that they produce higher harmonics. Not yet widely known is that subharmonics can also be produced by some nonlinear systems. These then seem to spontaneously divide the applied frequency f_0 to yield, for example, exactly half that frequency (or exactly one-third). This phenomenon has become known as a period-doubling (-tripling) bifurcation. A large class of systems has been found to show period doubling, among them driven nonlinear oscillators. A peculiar feature of the period-doubling bifurcation is that it occurs in sequences; that is, when one period-doubling bifurcation has occurred, it is likely that further period doubling will occur upon altering a parameter of the system, and so on, often in an infinite series. Acoustic cavitation has been one of the first experimental examples known to exhibit this series. In Fig. 2, the upper right-hand diagram shows the noise spectrum after further period doubling to $\frac{1}{4}f_0$. The doubling sequence can be observed via $\frac{1}{8}f_0$ and $\frac{1}{16}f_0$ up to $\frac{1}{32}f_0$ (not shown here). It is obvious that the spectrum is rapidly "filled" with lines and gets more and more dense. The limit of the infinite series yields an aperiodic motion, a densely packed power spectrum (not homogeneously), that is, broadband noise (but characteristically colored by lines). One such noise spectrum is shown in Fig. 2 (lower right-hand diagram). Thus, at least one way of turning

a pure tone into broadband noise has been found—via successive period doubling.

This finding has a deeper implication. If a system becomes aperiodic through the phenomenon of repeated period doubling, then this is a strong indication that the irregularity attained in this way is of simple deterministic origin. This implies that acoustic cavitation noise is not a basically statistical phenomenon but a deterministic one. It also implies that a description of the system with usual statistical means may not be appropriate and that a successful description by some deterministic theory may be feasible.

III. A FRACTAL NOISE ATTRACTOR

In Section II the sound signal has been treated by Fourier analysis. Fourier analysis is a decomposition of a signal into a sum of simple waves (normal modes) and is said to give the degrees of freedom of the described system. Chaos theory shows that this interpretation must be abandoned. Broadband noise, for instance, is usually thought to be due to a high (nearly infinite) number of degrees of freedom that superposed yield noise. Chaotic systems, however, have the ability to produce noise with only a few (nonlinear) degrees of freedom, that is, with only a few dependent variables. Also, it has been found that continuous systems with only three dependent variables are capable of chaotic motions and thus, producing noise. Chaos theory has developed new methods to cope with this problem. One of these is phase-space analysis, which in conjunction with fractal dimension estimation is capable of yielding the intrinsic degrees of freedom of the system. This method has been applied to inspect acoustic cavitation noise. The answer it may give is the dimension of the dynamical system producing acoustic cavitation noise. *See* SERIES.

The sampled noise data are first used to construct a noise attractor in a suitable phase space. Then the (fractal) dimension of the attractor is determined. The procedure to construct an attractor in a space of chosen dimension n simply consists in combining n samples (not necessarily consecutive ones) to an n-tuple, whose entries are interpreted as the coordinate values of a point in n-dimensional Euclidian space. An example of a noise attractor constructed in this way is given in Fig. 3. The attractor has been obtained from a time series of pressure values $\{p(kt_s);\ t = 1, \ldots, 4096;\ t_s = 1\ \mu\text{sec}\}$ taken at a sampling frequency of $f_s = 1/t_s = 1$ MHz by forming the three-tuples $[p(kt_s),\ p(kt_s + T), p(kt_s + 2T)], k = 1, \ldots, 4086$, with $T = 5\ \mu$sec. The frequency of the driving sound field has been 23.56 kHz. The attractor in Fig. 3 is shown from different views to demonstrate its nearly flat structure. It is most remarkable that not an unstructured cluster of points is obtained as is expected for noise, but a quite well-defined

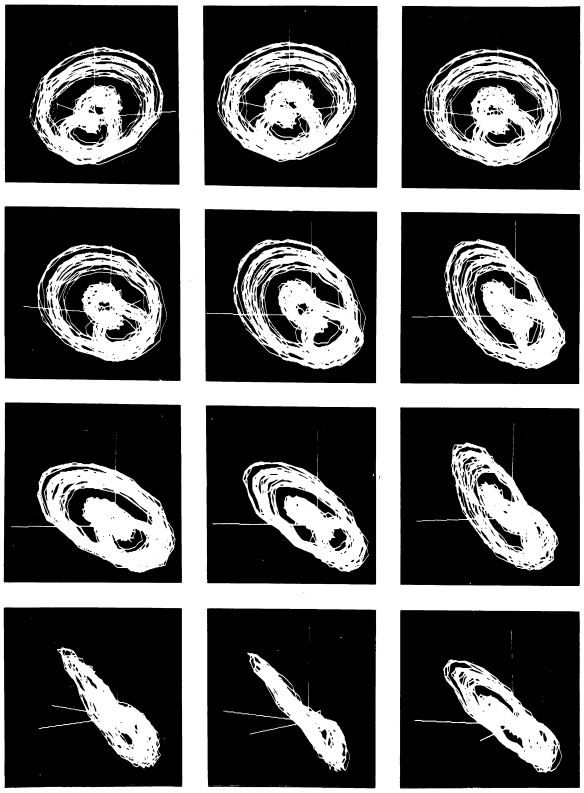


FIGURE 3 Strange attractor of acoustic cavitation noise obtained by phase–space analysis of experimental data (a time series of pressure values sampled at 1 MHz). The attractor is rotated to visualize its three-dimensional structure. (Courtesy of J. Holzfuss. From Lauterborn, W. (1986). *In* "Frontiers in Physical Acoustics" (D. Sette, ed.), pp. 124–144, North Holland, Amsterdam.)

object. This suggests that the dynamical system producing the noise has only a few nonlinear degrees of freedom. The flat appearance of the attractor in a three-dimensional phase space (Fig. 3) suggests that only three essential degrees are needed for the system. This is confirmed by a fractal dimension analysis, which yields a dimension of $d = 2.5$ for this attractor. Unfortunately, a method has not yet been conceived of how to construct the equations of motion from the data.

IV. LYAPUNOV ANALYSIS

Chaotic systems exhibit what is called sensitive dependence on initial conditions. This expression has been introduced to denote the property of a chaotic system that small differences in the initial conditions, however small, are persistently magnified because of the dynamics of the system. This property is captured mathematically by the notion of Lyapunov exponents and Lyapunov spectra. Their definition can be illustrated by the deformation of a small

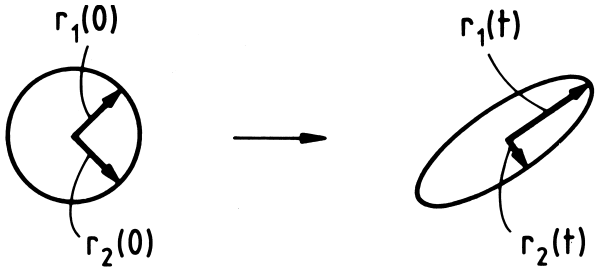


FIGURE 4 Idea for defining Lyapunov exponents. A small sphere in phase space is deformed to an ellipsoid, indicating expansion or contraction of neighboring trajectories.

sphere of initial conditions along a fiducial trajectory (see Fig. 4). The expansion or contraction is used to define the Lyapunov exponents $\lambda_i, i = 1, 2, \ldots, m$, where m is the dimension of the phase space of the system. When, on the average, for example, $r_1(t)$ is larger than $r_1(0)$, then $\lambda_1 > 0$ and there is a persistent magnification in the system. The set $\{\lambda_i, i = 1, \ldots, m\}$, whereby the λ_i usually are ordered $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, is called the Lyapunov spectrum.

In dissipative systems, the final motion takes place on attractors. Besides the fractal dimension, as discussed in the previous section, the Lyapunov spectrum may serve to characterize these attractors. When at least one Lyapunov exponent is greater than zero, the attractor is said to be chaotic. Progress in the field of nonlinear dynamics has made possible the calculation of the Lyapunov spectrum from a time series. It could be shown that acoustic cavitation in the region of broadband noise emission is characterized by one positive Lyapunov exponent.

V. PERIOD-DOUBLING BUBBLE OSCILLATIONS

Thus far, only the acoustic signal has been investigated. An optic inspection of the liquid inside the piezoelectric cylinder (see Fig. 1) reveals that a highly structured cloud of bubbles or cavities is present (Fig. 5) oscillating and moving in the sound field. It is obviously these bubbles that produce the noise. If this is the case, the bubbles must

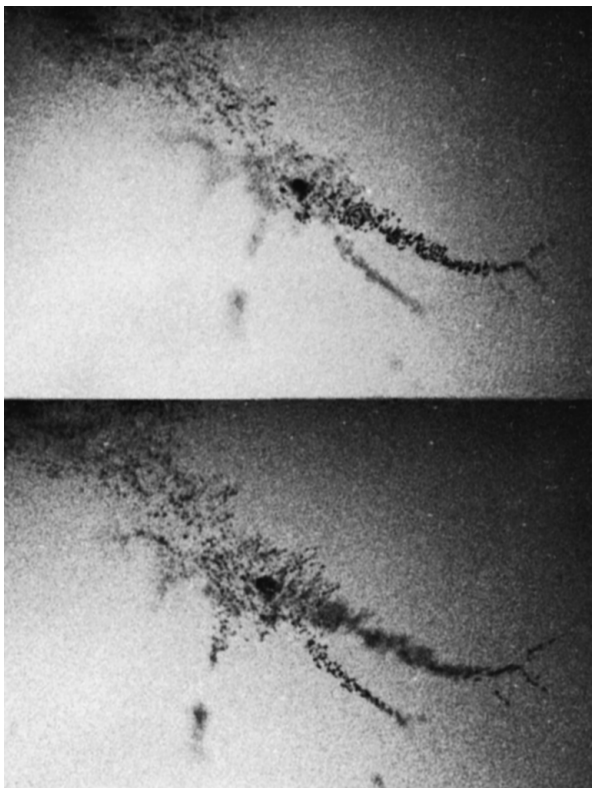


FIGURE 5 Acoustic cavitation bubble field in water inside a cylindrical piezoelectric transducer of about 7 cm in diameter. Two planes in depth are shown about 5 mm apart. The pictures are obtained by photographs from the reconstructed three-dimensional image of a hologram taken with a ruby laser.

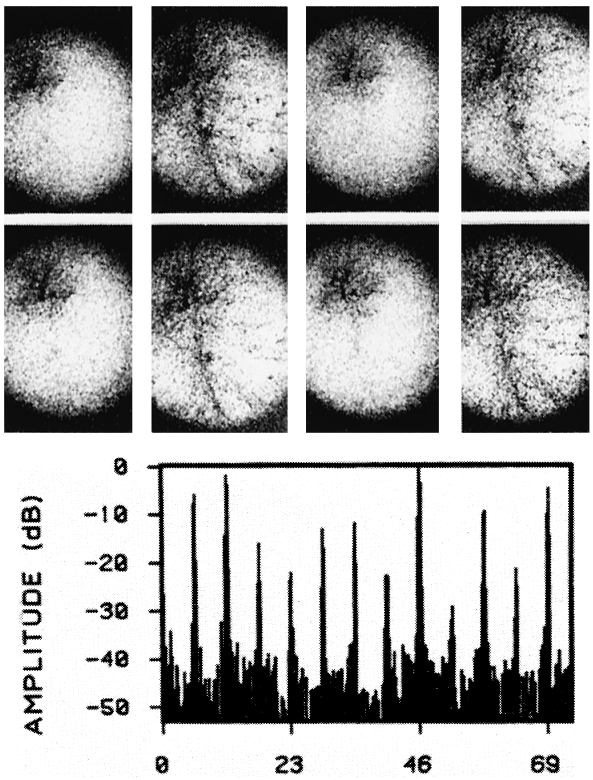


FIGURE 6 Reconstructed images from (a) a holographic series taken at 23.100 holograms per second of bubbles inside a piezoelectric cylinder driven at 23.100 Hz and (b) the corresponding power spectrum of the noise emitted. Two period-doublings have taken place.

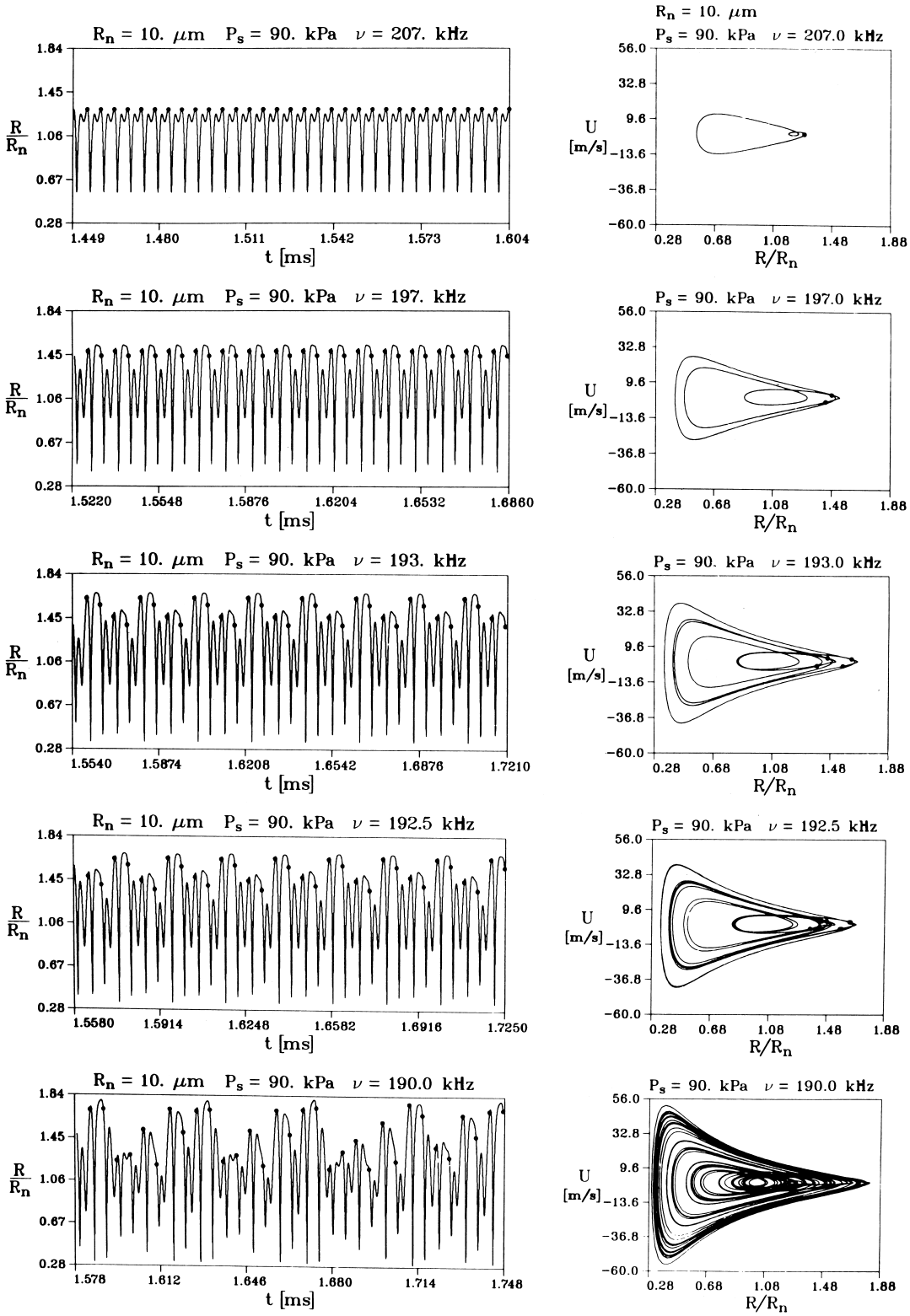


FIGURE 7 Period-doubling route to chaos for a driven bubble oscillator. Left column: radius-time solution curves; middle left column: trajectories in phase space; middle right column: Poincaré section plots: right column: power spectra. R_n is the radius of the bubble at rest, P_s and ν are the pressure amplitude and frequency of the driving sound field, respectively. (From Lauterborn, W., and Parlitz, U. (1988). *J. Acoust. Soc. Am.* **84,** 1975.)

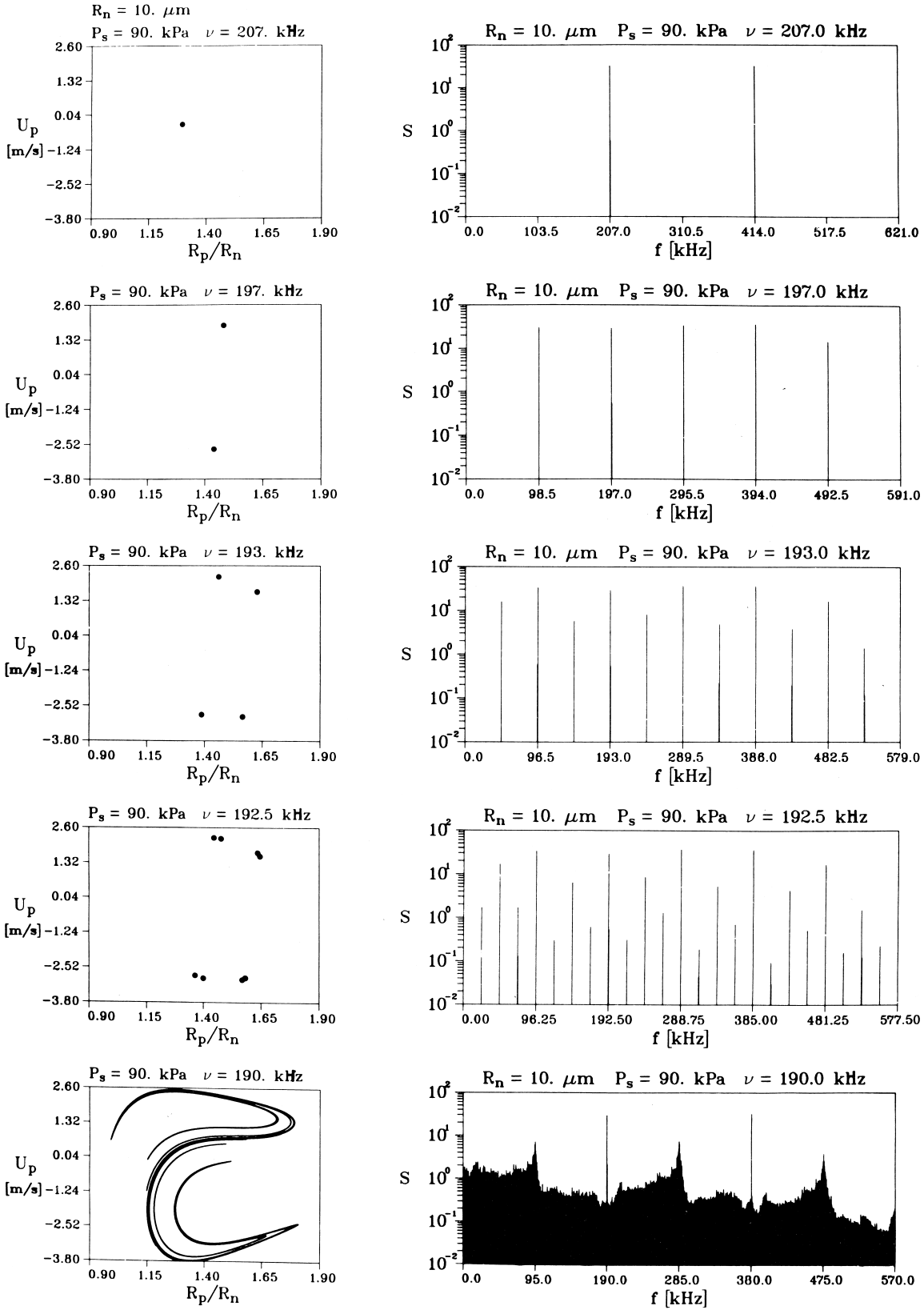


FIGURE 7 (*Continued*)

move chaotically and should show the period-doubling sequence encountered in the noise output. This has been confirmed by holographic investigations where once per period of the driving sound field a hologram of the bubble field has been taken. Holograms have been taken because the bubbles move in three dimensions, and it is difficult to photograph them at high resolution when an extended depth of view is needed. In one experiment the driving frequency was 23,100 Hz, which means 23,100 holograms per second have been taken. The total number of holograms, however, was limited to a few hundred. Figure 6a gives an example of a series of photographs taken from a holographic series. In this case, two period-doubling bifurcations have already taken place since the oscillations only repeat after four cycles of the driving sound wave. The first period doubling is strongly visible; the second one can only be seen by careful inspection. Figure 6b gives the noise power spectrum taken simultaneously with the holograms. The acoustic measurements show both period doublings more clearly than the optical measurement (documented in Fig. 6a) as the $\frac{1}{4} f_0$ ($f_0 =$ 23.1 kHz) spectral line is strongly present together with its harmonics.

VI. THEORY OF DRIVEN BUBBLES

A theory has not yet been developed that can account for the dynamics of a bubble field as shown in Fig. 5. The most advanced theory is only able to describe the motion of a single spherical bubble in a sound field. Even with suitable neglections the model is a highly nonlinear ordinary differential equation of second order for the radius R of the bubble as a function of time. With a sinusoidal driving term (sound wave) the phase space is three dimensional, just sufficient for a dynamical system to show irregular (chaotic) motion. The model is an example of a driven nonlinear oscillator for which chaotic solutions in certain parameter regions are by now standard. However, period doubling and irregular motion were found in the late 1960s in numerical calculations when chaos theory was not yet available and thus the interpretation of the results difficult. The surprising fact is that already this simple model of a purely spherically oscillating bubble set into oscillation by a sound wave yields successive period doubling up to chaotic oscillations. Figure 7 demonstrates the period-doubling route to chaos in four ways. The leftmost column gives the radius of the bubble in the sound field as a function of time, where the dot on the curve indicates the lapse of a full period of the driving sound field. The next column shows the corresponding trajectories in the plane spanned by the radius of the bubble and its velocity. The dots again mark the lapse of a full period of the driving sound field. The third column shows so-called Poincaré section plots. Here, only the dots after the lapse of one full period of the driving sound field are plotted in the radius–velocity plane of the bubble motion. Period doubling is seen most easily here and also the evolution of a strange (or chaotic) attractor. The rightmost column gives the power spectra of the radial bubble motion. The filling of the spectrum with successive lines in between the old lines is evident, as is the ultimate filling when the chaotic motion is reached.

A compact way to show the period-doubling route to chaos is by plotting the radius of the bubble as a function of a parameter of the system that can be varied, e.g., the frequency of the driving sound field. Figure 8a gives an example for a bubble of radius at rest of $R_n = 10\ \mu$m, driven by a sound field of frequency ν between 390 kHz and 510 kHz at a pressure amplitude of $P_s = 290$ kPa. The period-doubling cascade to chaos is clearly visible. In the chaotic region, "windows" of periodicity show

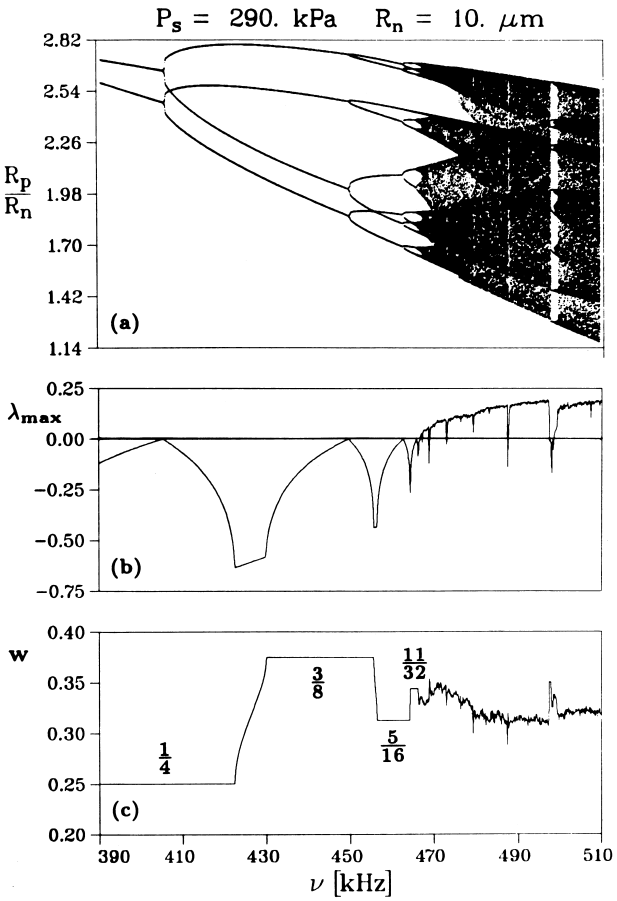


FIGURE 8 (a) A period-doubling cascade as seen in the bifurcation diagram. (b) The corresponding largest Lyapunov exponent λ_{max}. (c) The winding number w. (From Parlitz, U. *et al.* (1990). *J. Acoust. Soc. Am.* **88,** 1061.)

up as regularly experienced with other chaotic systems. In Fig. 8b the largest Lyapunov exponent λ_{max} is plotted. It is seen that $\lambda_{max} > 0$ when the chaotic region is reached. Figure 8c gives a further characterization of the system by the winding number w. The winding number describes the winding of a neighboring trajectory around the given one per period of the bubble oscillation. It can be seen that this quantity changes quite regularly in the period-doubling sequence, and rules can be given for this change.

The driven bubble system shows resonances at various frequencies that can be labeled by the ratio of the linear resonance frequency of the bubble to the driving frequency of the sound wave. Figure 9 gives an example of the complicated response characteristic of a driven bubble. At somewhat higher driving than given in the figure the oscillations start to become chaotic. A chaotic bubble attractor is shown in Fig. 10. To better reveal its structure, it is not the total trajectory that is plotted but only the points in the velocity–radius plane of the bubble wall at a fixed phase of the driving. These points hop around on the attractor in an irregular fashion. These chaotic bubble oscillations must be considered as the source of the chaotic sound output observed in acoustic cavitation.

VII. OTHER SYSTEMS

Are there other systems in acoustics with chaotic dynamics? The answer is surely yes, although the subtleties of chaotic dynamics make it difficult to easily locate them.

When looking for chaotic acoustic systems, the question arises as to what ingredients an oscillatory system, as an acoustic one, must possess to be susceptible to chaos. The full answer is not yet known, but some understanding is emerging. A necessary, but unfortunately not sufficient, ingredient is nonlinearity. Next, period doubling is known to be a precursor of chaos. It is a peculiar fact that, when one period doubling has occurred, another one is likely to appear, and indeed a whole series with slight alterations of parameters. Further, the appearance of oscillations when a parameter is altered points to an intrinsic instability of a system and thus to the possibility of becoming a chaotic one. After all, two distinct classes can be formulated: (1) periodically driven passive nonlinear systems (oscillators) and (2) self-excited systems (oscillators). Passive means that in the absence of any external driving the system stays at rest as, for instance, a pendulum does. But a pendulum has the potential to oscillate chaotically when being driven periodically, for instance by a sinusoidally

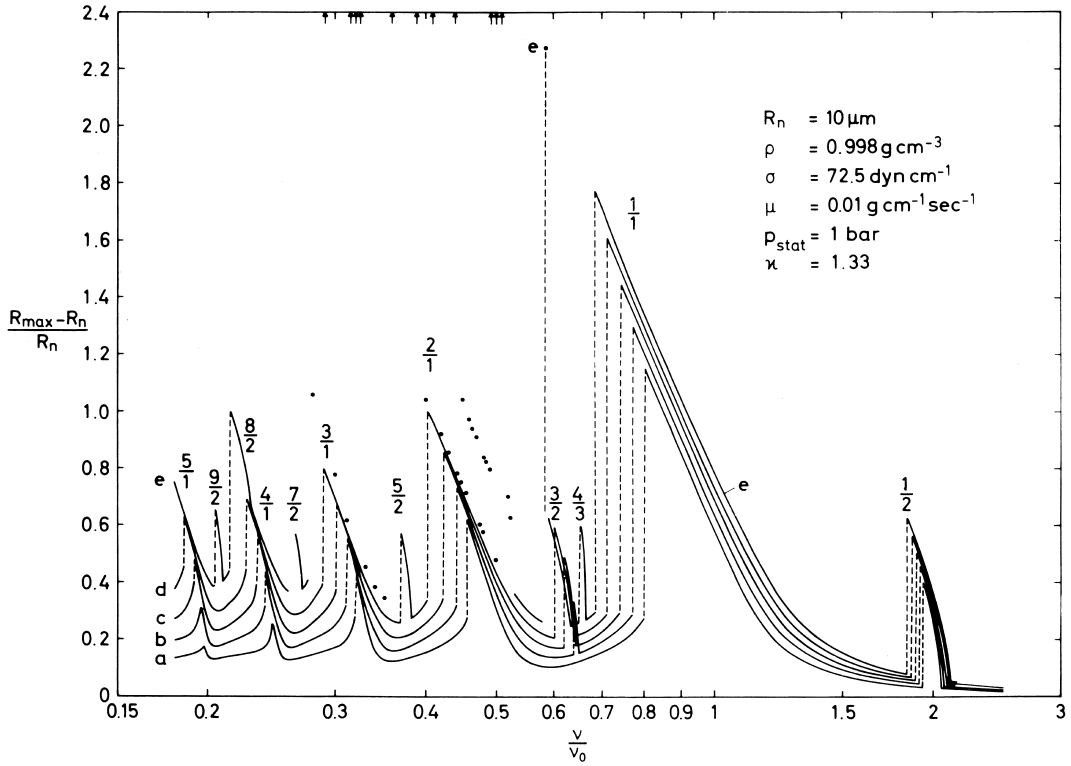


FIGURE 9 Frequency response curves (resonance curves) for a bubble in water with a radius at rest of $R_n = 10$ μm for different sound pressure amplitudes p_A of 0.4, 0.5, 0.6, 0.7, and 0.8 bar. (From Lauterborn, W. (1976). *J. Acoust. Soc. Am.* **59,** 283.)

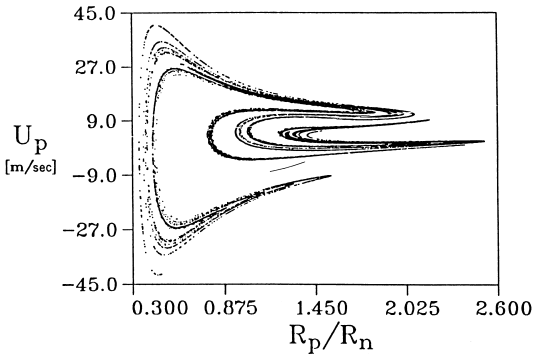


FIGURE 10 A numerically calculated strange bubble attractor ($P_s = 300$ kPa, $\nu = 600$ kHz). (Courtesy of U. Parlitz.)

varying torque. This is easily shown experimentally by the repeated period doubling that soon appears at higher periodic driving. Self-excited systems develop sustained oscillations from seemingly constant exterior conditions. One example is the Rayleigh-Bénard convection, where a liquid layer is heated from below in a gravitational field. The system goes chaotic at a high enough temperature difference between the bottom and surface of the liquid layer. Self-excited systems may also be driven, giving an important subclass of this type. The simplest model in this class is the driven van der Pol oscillator. A real physical system of this category is the weather (the atmosphere). It is periodically driven by solar radiation with the low period of 24 hr, and it is a self-excited system, as already constant heating by the sun may lead to Rayleigh-Bénard convection as observed on a faster time scale.

The first reported period-doubled oscillation from a periodically driven passive system dates back to Faraday in 1831. Starting with the investigation of sound-emitting, vibrating surfaces with the help of Chladni figures, Faraday used water instead of sand, resulting in vibrating a layer of liquid vertically. He was very astonished about the result: regular spatial patterns of a different kinds appeared and, above all, these patterns were oscillating at half the frequency of the vertical motion of the plate. Photography was not yet invented to catch the motion, but Faraday may well have seen chaotic motion without knowing it. It is interesting to note that there is a connection to the oscillation of bubbles as considered before. Besides purely spherical oscillations, bubbles are susceptible to surface oscillations as are drops of liquid. The Faraday case of a vibrating flat surface of a liquid may be considered as the limiting case of either a bubble of larger and larger size or a drop of larger and larger size, when the surface is bent around up or down. Today, the Faraday patterns and Faraday oscillations can be observed better, albeit still with difficulties as it is a three-dimensional (space), nonlinear, dynamical (time) system; that is, it requires three space coordinates and one time coordinate to be followed. This is at the border of present-day technology both numerically and experimentally. The latest measurements have singled out mode competition as the mechanism underlying the complex dynamics. Figure 11 gives two examples of oscillatory patterns: a periodic hexagonal structure (Fig. 11a) and

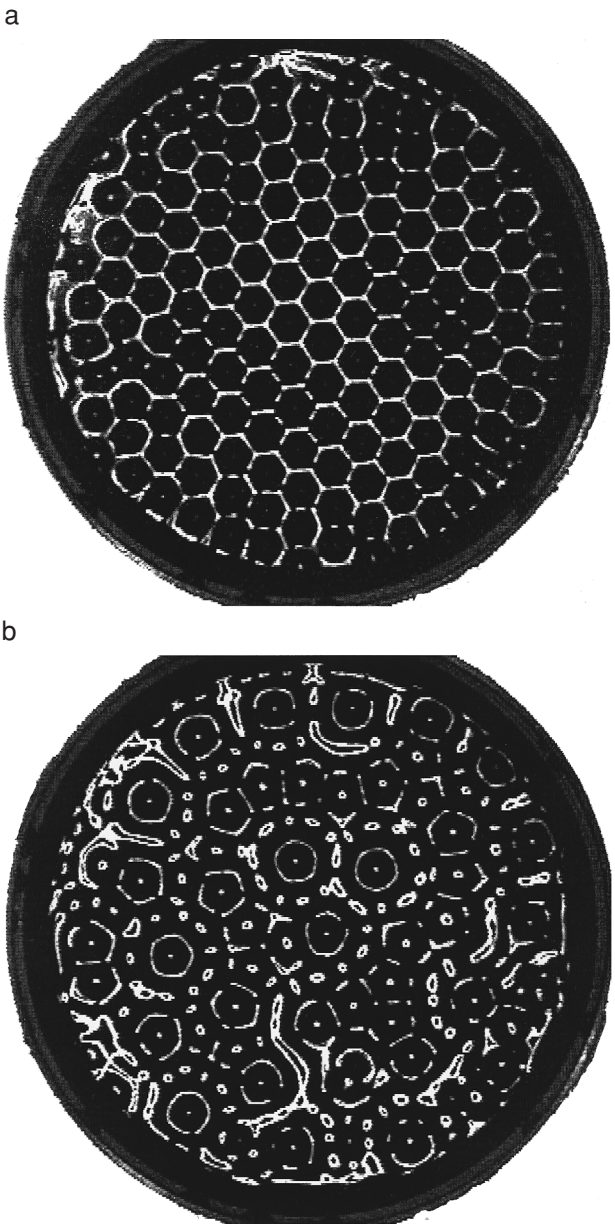


FIGURE 11 Two patterns appearing on the surface of a liquid layer vibrated vertically in a cylindrical container: (a) regular hexagonal pattern at low amplitude, and (b) pattern when approaching chaotic vibration. (Courtesy of Ch. Merkwirth.)

its dissolution on the way to chaotic motion (Fig. 11b) at the higher vertical driving oscillation amplitude of a thin liquid layer.

The other class of self-excited systems in acoustics is quite large. It comprises (1) musical instruments, (2) thermoacoustic oscillators as used today for cooling with sound waves, and (3) speech production via the vocal folds. Period doubling could be observed in most of these systems; however, very few investigations have been done so far concerning their chaotic properties.

VIII. PHILOSOPHICAL IMPLICATIONS

The results of chaos physics have shed new light on the relation between determinism and predictability and on how seemingly random (irregular) motion is produced. It has been found that deterministic laws do not imply predictability. The reason is that there are deterministic laws which persistently show a sensitive dependence on initial conditions. This means that in a finite, mostly short time any significant digit of a measurement has been lost, and another measurement after that time yields a value that appears to come from a random process. Chaos physics has thus shown a way of how random (seemingly random, one must say) motion is produced out of determinism and has developed convincing methods (some of them exemplified in the preceding sections on acoustic chaos) to classify such motion. Random motion is thereby replaced by chaotic motion. Chaos physics suggests that one should not resort too quickly to statistical methods when faced with irregular data but instead should try a deterministic approach. Thus, chaos physics has sharpened our view considerably on how nature operates.

But, as always in physics, when progress has been made on one problem other problems pile up. Quantum mechanics is thought to be the correct theory to describe nature. It contains "true" randomness. But, what then about the relationship between classical deterministic physics and quantum mechanics? Chaos physics has revived interest in these questions and formulated new specific ones, for instance, on how chaotic motion crosses the border to quantum mechanics. What is the quantum mechanical equivalent to sensitive dependence on initial conditions?

The exploration of chaos physics, including its relation to quantum mechanics, is therefore thought to be one of the big scientific enterprises of the new century. It is hoped that acoustic chaos will accompany this enterprise further as an experimental testing ground.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • CHAOS • FOURIER SERIES • FRACTALS • QUANTUM MECHANICS

BIBLIOGRAPHY

Lauterborn, W., and Holzfuss, J. (1991). "Acoustic chaos." *Int. J. Bifurcation and Chaos* **1**, 13–26.

Lauterborn, W., and Parlitz, U. (1988). Methods of chaos physics and their application to acoustics. *J. Acoust. Soc. Am.* **84,** 1975–1993.

Parlitz, U., Englisch, V., Scheffezyk, C., and Lauterborn, W. (1990). "Bifurcation structure of bubble oscillators." *J. Acoust. Soc. Am.* **88,** 1061–1077.

Ruelle, D. (1991). "Chance and Chaos," Princeton Univ. Press, Princeton, NJ.

Schuster, H. G. (1995). "Deterministic Chaos: An Introduction," Wiley-VCH, Weinheim.



Acoustical Measurement

Allan J. Zuckerwar
NASA Langley Research Center

I. Instruments for Measuring the Properties of Sound
II. Instruments for Processing Acoustical Data
III. Examples of Acoustical Measurements

GLOSSARY

Anechoic Having no reflections or echoes.
Audio Pertaining to sound within the frequency range of human hearing, nominally 20 Hz to 20 kHz.
Coupler Small leak-tight enclosure into which acoustic devices are inserted for the purpose of calibration, measurement, or testing.
Diffuse field Region of uniform acoustic energy density.
Free field Region where sound propagation is unaffected by boundaries.
Harmonic Pertaining to a pure tone, that is, a sinusoidal wave at a single frequency: an integral multiple of a fundamental tone.
Infrasonic Pertaining to sound at frequencies below the limit of human hearing, nominally 20 Hz.
Reverberant Highly reflecting.
Ultrasonic Pertaining to sound at frequencies above the limit of human hearing, nominally 20 kHz.

A SOUND WAVE propagating through a medium produces deviations in pressure and density about their mean or static values. The deviation in pressure is called the acoustic or sound pressure, which has standard international (SI) units of pascal (Pa) or newton per square meter (N/m^2). Because of the vast range of amplitude covered in acoustic measurements, the sound pressure is conveniently represented on a logarithmic scale as the sound pressure level (SPL). The SPL unit is the decibel (dB), defined as

$$\mathrm{SPL(dB)} = 20 \log(p/p_0)$$

in which p is the root mean square (rms) sound pressure amplitude and p_0 the reference pressure of 20×10^{-6} Pa. The equivalent SPLs of some common units are the following:

pascal (Pa)	93.98 dB	psi (lb/in.^{2})	170.75 dB
atmosphere (atm)	194.09	torr (mm Hg)	136.48
bar	193.98	dyne/cm^2	73.98

The levels of some familiar sound sources and environments are listed in Table I.

The displacement per unit time of a fluid particle due to the sound wave, superimposed on that due to its thermal motion, is called the acoustic particle velocity, in units of meters per second. Determination of the sound pressure and acoustic particle velocity at every point completely specifies an acoustic field, just as the voltages and currents completely specify an electrical network. Thus, acoustical instrumentation serves to measure one of these quantities or both. Since in most cases the relationship between

TABLE I Representative Sound Pressure Levels of Familiar Sound Sources and Environments

Source or environment	Level (dB)
Concentrated sources: re 1 m	
Four-jet airliner	155
Pipe organ, loudest	125
Auto horn, loud	115
Power lawnmower	100
Conversation	60
Whisper	20
Diffuse environments	
Concert hall, loud orchestra	105
Subway interior	95
Street corner, average traffic	80
Business office	60
Library	40
Bedroom at night	30
Threshold levels	
Of pain	130
Of hearing impairment, continous exposure	90
Of hearing	0
Of detection, good microphone	−2

sound pressure and particle velocity is known, it is sufficient to measure only one quantity, usually the sound pressure. The scope of this article is to describe instrumentation for measuring the properties of sound in fluids, primarily in air and water, and in the audio (20 Hz–20 kHz) and infrasonic (<20 Hz) frequency ranges. Although many instrumentation techniques conform to national and international standards, the standards are not cited in the text, but a selected list is given in the bibliography.

I. INSTRUMENTS FOR MEASURING THE PROPERTIES OF SOUND

A. Measurement of Sound Pressure

A device that senses a sound pressure in a gas and provides a proportional electrical output voltage is a microphone. Functionally, microphones fall into two categories: entertainment or broadcasting microphones and measurement microphones. Entertainment microphones, comprising mainly electret, ribbon, and moving-coil microphones, conform to the requirements of speech and music and have preferred directionality. Measurement microphones, on the other hand, may have capabilities that extend well beyond these requirements, both in frequency response and in dynamic range, and are for the most part omnidirectional. Nevertheless, the distinguishing mark of a measurement microphone is consistent performance in the face of prolonged service, exposure to various environmental conditions, and the passage of time. Recalibration often yields microphone sensitivities that are repeatable to within tenths of a decibel—a feature not required in entertainment microphones. The common types of measurement microphone are air condenser, electret condenser, ceramic, and piezoresistive microphones.

1. Air Condenser Microphone

Beginning with the successful operational unit reported by Wente in 1917, the evolution of condenser microphone design culminated in the celebrated Western Electric model 640 AA in 1948, which serves as the prototype of modern air condenser microphones. Because its operation depends on a purely geometric effect, the air condenser microphone remains the most stable and most widely used measurement microphone today. Its basic construction is shown in Fig. 1. An incident sound pressure p excites motion of the membrane, changing the capacitance between the membrane and backplate and producing a proportional output voltage. The mechanical and electrical functions will be described separately in that order.

As the membrane vibrates, it compresses and expands the air layer in the gap and creates a reaction pressure, which opposes motion of the membrane. The reaction pressure is partially relieved by the flow of air through the openings in the backplate, and these determine the damping of the membrane–air layer system. The backplate may contain one or more "rings" of holes and nearly always a slot around its periphery. The flow of air through the openings depends on the pressure difference across them. Because the pressure at any one opening depends on the pressures at all the other openings, both in the gap and in the backchamber, the pressures at the openings are coupled together at both locations and their analysis is subject to a very complicated boundary–value problem. However, through simplifying assumptions the problem can be solved approximately but accurately, taking into account the specifics of the backplate configuration. The solution yields a mechanical sensitivity M_m (m/Pa) of the form:

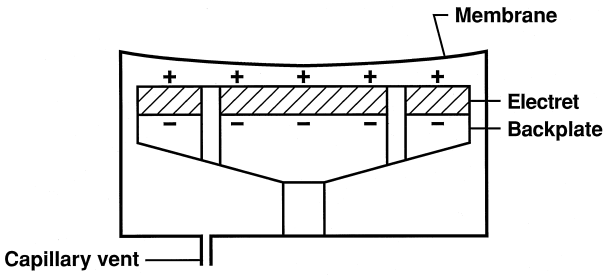


FIGURE 1 Basic construction of an air condenser microphone.

$$M_{\mathrm{m}} \equiv \frac{d}{p} = \frac{1}{TK^2} \cdot \frac{J_2(Ka)}{J_0(Ka) + D} \tag{1}$$

where d is the mean membrane displacement from equilibrium (m), p the incident sound pressure (N/m^2), K the wave number for sound propagation in the membrane (m^{-1}), a the membrane radius (m) and T its tension (N/m), and the J's are the Bessel functions of the first kind. The complex term D, which accounts for the effect of the reaction pressure in the gap, can be expressed explicitly in terms of the air layer and backplate parameters, but such a derivation is beyond the scope of this article. The membrane wave number is given by:

$$K = 4\pi^2 f(\rho_{\mathrm{M}} t_{\mathrm{M}}/T)^{1/2} \tag{2}$$

where ρ_{M} is the membrane density (kg/m^3), t_{M} the membrane thickness (m), and f the acoustic frequency (Hz). The upper cutoff frequency of the microphone lies close to the undamped fundamental resonant frequency of the membrane:

$$f_{\mathrm{R}} = \left(T/6.285\rho_{\mathrm{M}} t_{\mathrm{M}} a^2\right)^{1/2} \tag{3}$$

At frequencies well below f_{R}, that is, over the normal operating range of the microphone, $Ka \ll 1$ and the Bessel functions in Eq. (1) can be represented by their leading terms. The microphone can be represented by the equivalent lumped elements, the first four terms in the expansion of Eq. (1), shown in Fig. 2a. Its acoustic impedance (Section I.E) is

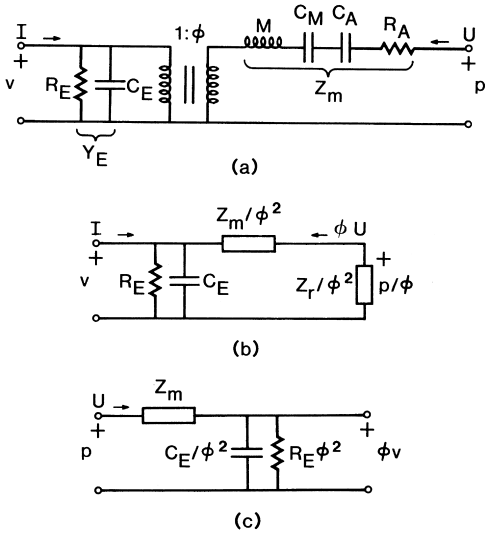


FIGURE 2a (a) Equivalent lumped element representation of a condenser microphone. (b) As a transmitter with the mechanical elements referred to the electrical side. (c) As a receiver with the electrical elements referred to the mechanical side. The transformation ratio ϕ is equal to the static charge on the backplate (or membrane) divided by the volume of the gap and microphone admittance.

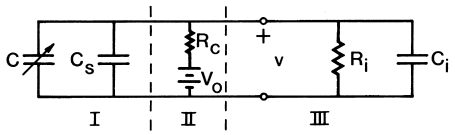


FIGURE 2b Polarization circuit of an air condenser microphone, consisting of (I) the condenser microphone, (II) the charging network, and (III) the input elements of the preamplifier.

$$Z_{\mathrm{m}} \equiv \frac{p}{U} = \frac{p}{j\omega d\pi a^2} = \frac{1}{j\omega M_{\mathrm{m}}\pi a^2}$$
$$= j\omega M + \frac{1}{j\omega}\left(\frac{1}{C_{\mathrm{M}}} + \frac{1}{C_{\mathrm{A}}}\right) + R_{\mathrm{A}} \tag{4}$$

where ω is the angular frequency $= 2\pi f$ and U the volume velocity of the incident sound. In terms of the microphone parameters, the membrane mass M, membrane compliance C_{M}, air layer compliance C_{A}, and air layer resistance R_{A} are

$$M = \tfrac{4}{3}\left(\rho_{\mathrm{M}} t_{\mathrm{M}}/\pi a^2\right) \tag{5}$$

$$C_{\mathrm{M}} = (\pi a^2)^2/8\pi T \tag{6}$$

$$C_{\mathrm{A}} = (\pi a^2)^2/8T D' \tag{7}$$

$$R_{\mathrm{A}} = 8\pi T D''/\omega(\pi a^2)^2 \tag{8}$$

in which D' and D'' are the real and imaginary parts of D. Typical values for the $\frac{1}{2}$- and 1-in. microphones are listed in Table II.

Most microphones are designed such that the low-frequency membrane displacement is controlled by the membrane compliance. In this case, the mechanical sensitivity can be approximated by

$$M_{\mathrm{m}} \approx a^2/8T \tag{9}$$

The membrane motion can be used to provide a proportion voltage by the arrangement shown in Fig. 2b. The electrical circuit is divided into three sections, representing (I) the microphone, having a time-varying membrane–backplate capacitance C, and stray capacitance C_{s}; (II) a

TABLE II Representative Values of the Lumped Elements of $\frac{1}{2}$- and 1-in. Condenser Microphones at Midband Frequencies and 1 atm Pressure

Lumed element	$\frac{1}{2}$-in.	1-in.	
M	950	300	kg/m^4
C_{M}	5	100	$\times 10^{-14}$ m^5/N
C_{A}	90	500	$\times 10^{-14}$ m^5/N
R_{A}	15	2	$\times 10^7 \cdot$ sec/m^5

charging network, consisting of a polarization voltage V_0 and charging resistance R_c; and (III) the input resistance R_i and capacitance C_i of a preamplifier (once called a cathode follower). The preamplifier is placed as close to the microphone cartridge as possible in order to minimize the input capacitance C_i. A blocking capacitor before the preamplifier is not shown. The polarization voltage source maintains a static charge on the capacitor. At frequencies above a certain lower limiting frequency, dependent on the charging time constant of the circuit (see Section I.A.9), the electrical sensitivity M_e can be approximated by:

$$M_e \equiv \frac{\upsilon}{d} = \frac{1}{d}\frac{\Delta C}{C_E}V_0 = \frac{V_0}{d_0} \tag{10}$$

where ΔC is the variation in C, $C_E = C + C_s + C_i$, and d_0 is the static gap between the microphone and backplate.

The overall microphone sensitivity is the product of Eqs. (9) and (10):

$$M_p = M_m M_e \approx a^2 V_0 / 8T d_0 \tag{11}$$

Despite its approximate nature, Eq. (11) illustrates, together with Eq. (3), the effect of design parameters on microphone performance. A high sensitivity is favored by a large membrane radius, high polarization voltage, low membrane tension, and small membrane–backplate gap. A high-frequency response is favored by a high membrane tension and small membrane density, thickness, and radius. Choice of the ratio a^2/T, which plays conflicting roles regarding sensitivity and frequency response, requires a design compromise. Generally, the tension is made as high as practical and the frequency response is controlled by the membrane radius a.

A good membrane material has high tensile strength in order to maintain high tension, high ductility so that it can be stretched tightly without cracking or wrinkling, and good resistance to corrosion. Some suitable materials are finegrained nickle, titanium, and certain grades of stainless steel. Typical values of thickness and tension are $t_M = 5$ μm and $T = 2000$–4000 N/m, corresponding to a tensile stress of 4–8×10^8 N/m^2. In some microphones, the tension can be adjusted by means of a tightening ring, controlled by the turn of a screw after the membrane is clamped or welded in place. In practice, the tension is adjusted beyond the design value and then reduced by heat treatment for additional stability. The membrane is exceedingly delicate and usually covered with a protective grid.

The polarization voltage and static gap are determined by limitations on the electric field in the gap and by practical electronic considerations. Typical values are $V_0 = 200$ V and $d_0 = 20$ μm. Because the polarization voltage reduces d_0 due to electrostatic attraction, considerations of electrical and mechanical stability may be important. In battery-operated units, a polarization voltage of $V_0 = 28$ V is commonly used. The electrical resistance $R_E = R_c \| R_i$ is of the order of gigaohms, corresponding to a charging time constant of several tenths of a second.

2. Electret Condenser Microphone

An electret material possesses a permanent electrical dipole moment. When used in a condenser microphone, it provides the polarization voltage between the membrane and backplate in place of the external supply voltage. Electret materials are generally high-resistivity polymers, a prime example being PTFE Teflon. They are fabricated by heating a film of the material almost to its melting point and subjecting it to an intense electric field. The net dipole moment results from either rotation of permanent dipoles in polar materials or from migration of free charge carriers. In either case, when the material is cooled to room temperature the net dipole moment is "frozen-in."

A typical construction is shown in Fig. 3. Here the electret is bonded to the backplate. This has an advantage over arrangements where the electret is bonded to the membrane because electret materials are not very suitable for performing the mechanical function of a membrane. The lower surface makes electrical contact with the backplate and thus is at the same potential as the membrane (a metal). The voltage V_0 at the upper surface and across the gap is

$$V_0 = \sigma t / \varepsilon \varepsilon_0 \tag{12}$$

where σ is the surface charge density (C/m^2), t the foil thickness (m), ε the dielectric constant, and ε_0 the dielectric permittivity of free space (8.85×10^{-12} F/m, or farad per meter). Typical values of $\sigma = 10^{-4}$ C/m^2, $t = 2 \times 10^{-5}$ m, and $\varepsilon = 2$ lead to a voltage $V_0 \approx 100$ V, which is comparable to the polarization voltages used in air condenser microphones.

The capacity of an electret material to retain its surface charge is highly temperature dependent. In one test, an electret microphone stored at 50°C and 95% relative humidity lost sensitivity at a rate of ~1 dB/year. Under normal ambient conditions an electret microphone can be expected to retain its initial sensitivity for many years.

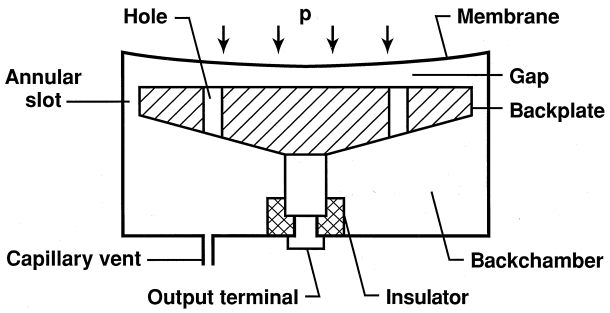


FIGURE 3 Typical construction of an air electret microphone. Sometimes the electret is bonded to the membrane.

The acoustic performance of the "backelectret" microphone is not much different from that of the air condenser microphone. The elimination of the external polarization voltage supply, however, has a significant advantage. The generation of a high-dc voltage and the extensive filtering needed to obtain a low noise floor, ripple, and hum require bulky components (except for battery-operated equipment). The absence of this requirement greatly enhances the miniaturization potential of electret-based instrumentation.

3. Ceramic Microphone

A ceramic microphone utilizes the piezoelectric effect, the generation of a surface charge density as the result of an applied stress. Traditionally, the sensing element is constructed of a piezoelectric ceramic, such as lead zirconate titanate (PZT).

The great rigidity of piezoelectric ceramics makes their fabrication in the form of a membrane impractical: rather, such elements would operate as vibrating plates, for which the microphone compliance is controlled not by static tension but by the elastic modulus. This would lead to two fundamental difficulties. First, the plate compliance would be too small to permit a reasonable mechanical sensitivity. Second, the sharp mechanical resonances of ceramic materials make them difficult to dampen.

An arrangement to circumvent these difficulties is shown in Fig. 4. The sound pressure on the membrane is transmitted to the ceramic element through a connecting rod. The sole purpose of the backplate is to provide the required mechanical damping of the membrane. Ceramic microphones are characterized by ruggedness, low cost, and simple electronics. However, there has been a trend to replace them with lowcost electrets.

4. Piezoresistive Microphone

The piezoresistive microphone exploits the physical effect known as "piezoresistivity"—the dependence of electrical resistivity upon mechanical stress or strain. The microphone membrane is a thin, micromachined silicon wafer on which dopant is diffused or implanted to form the resistors of a Wheatstone bridge, as shown in Fig. 5. Acoustical excitation deflects the membrane to generate a time-varying stress in the strategically positioned resistors. A proportional output voltage appears at the output of the bridge. Advantages include small size and low output impedance, making possible remote-control electronics and thus installation in limited confines. Representative specifications are given in Table III. Suited for high acoustic sound pressures, it is a favorite for wind tunnel and aerospace testing. A disadvantage is the temperature sensitivity of the resistors, requiring rather sophisticated compensation techniques.

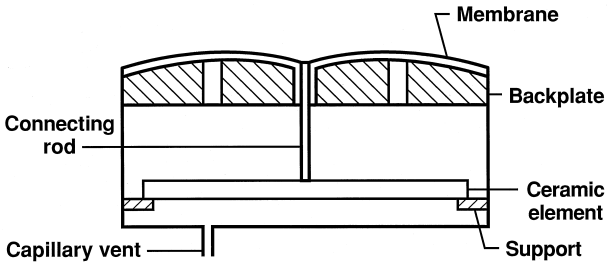


FIGURE 4 Basic construction of a ceramic microphone. The ceramic is usually a bimorph element, that is, two crystals sandwiched together to form a single assembly.

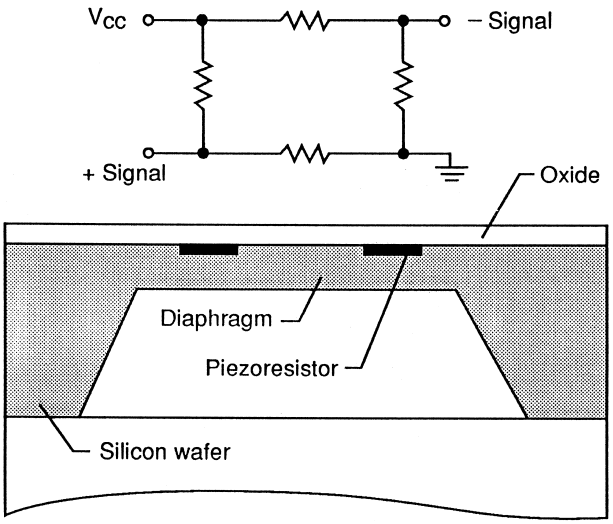


FIGURE 5 Basic construction of a piezoresistive microphone. Dopant is diffused or ion-implanted to form the piezoresistors.

5. Microphone Specifications

A microphone user examining a specification sheet will usually find the items listed in Table III. It is instructive to examine their meanings and implications. Specifications regarding environmental conditions will be considered later.

The *nominal size* refers to the outside diameter of the cartridge, which is slightly larger than the active membrane diameter.

The *open-circuit sensitivity* is the output voltage per unit sound pressure without the loading of the preamplifier input impedance. Sometimes this specification is given in decibels relative to a sensitivity of 1 V/Pa. For example, an open-circuit sensitivity of −40 dB is equivalent to 10 mV/Pa.

In a strict sense, the membrane *resonant frequency* is the fundamental resonant frequency of the membrane in

TABLE III Representative Microphone Specifications

Specifications	Microphone type					
	Air condenser			Electret	Ceramic	Piezoresistive
Nominal size (in.)	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1	0.092
Open circuit sensitivity (mV/Pa)	50	15	2	10	10	0.025
Resonant frequency (Hz)	8000	25,000	75,000	14,000	—	70,000
Frequency range, −2 dB (Hz)	2–7000	4–20,000	8–70,000	4–20,000	2–12,000	0–20,000
Dynamic range (dB)	15–145	25–160	35–170	30–145	25–150	80–190
Polarized capacitance (pf)	60	20	6	30	400	—
Equivalent air volume (cm^3)	0.15	0.01	0.0005	0.015	0.5	—

vacuum. In practical terms, it is the frequency at which the membrane displacement in air lags the applied pressure by 90°, which is less than the vacuum resonant frequency because of damping by the air layer.

At the *upper cutoff frequency*, the microphone sensitivity falls 2 dB, sometimes specified as 3 dB, on the high-frequency side of the damped membrane response. The specification is shown for pressure microphones and is different for free-field microphones (see Section I.A.6). The *lower cutoff frequency* is discussed in Section I.A.9.

The *dynamic range* is the range of sound pressure amplitudes over which the microphone operates linearly, usually specified in decibels. The upper limit is determined by total harmonic distortion, typically taken at 3 or 4%. In a system using a polarization voltage, the harmonic distortion is of electrical origin and not mechanical, that is, not due to nonlinear membrane displacement. This may not be true for systems using carrier electronics (see Section I.A.9). For a given sound pressure the harmonic distortion is less for small microphones because of lower output voltage. The lower limit of the dynamic range is determined by the noise floor, the rms output voltage over a specified frequency range (usually the "A-weighted" band, Section II.A.3) in the absence of sound. Since acoustic applications generally require a signal-to-noise ratio of better than 1 : 1 (0 dB), the lower limit of the dynamic range is usually specified as 5 dB above the noise floor. The two most important sources of noise are (1) Brownian motion of air molecules impinging on the membrane, primarily on the air layer side; and (2) noise generated in the preamplifier. The availability of high-quality field-effect transistors has reduced the preamplifier noise to a secondary role. Today it is possible to produce 1-in. condenser microphone systems having a noise floor several decibels below the threshold of hearing (0 dB) over the audible range of frequencies. It is interesting that the membrane displacement of a 1-in. condenser microphone having a 200-V polarization voltage and responding to a sound pressure of 0 dB is 10^{-13} m.

The *polarized capacitance* is the membrane–backplate capacitance with the polarization voltage applied. This is an important parameter in the determination of sensitivity and low-frequency response. A high capacitance helps reduce the noise floor.

Below the membrane resonant frequency the membrane acoustic impedance appears compliant, that is, dominated by C_M and C_A in Fig. 2. The *equivalent air volume* V_e is the volume of an enclosure that would present the same acoustic impedance as the membrane,

$$V_e = \gamma P_0 / j\omega Z_m = \gamma P_0 C_M \tag{13}$$

where γ is the specific heat ratio for air (=1.4), P_0 the ambient pressure, and Z_m the acoustic impedance of the microphone (see Fig. 2). If a microphone cartridge is inserted into a coupler, as for calibration purposes, then the equivalent volume must be added to the volume of the coupler.

6. Directional Properties

A microphone will not disturb a sound field if its dimensions are much smaller than the wavelength of the incident sound. For this reason the preamplifier (or adapter) is made as compact as possible, having a diameter not exceeding that of the microphone cartridge. Since the pressure distribution is uniform over the membrane area, the response of a microphone to this type of excitation is called the pressure response. An example is shown in Fig. 6. In this case, the mechanical damping of the microphone is designed for maximum flatness, corresponding to a mechanical quality factor of ∼0.7, and the microphone is called a pressure microphone. Most calibration methods provide a uniform pressure distribution and thus yield the pressure response.

As the wavelength in a free sound field approaches the dimensions of the microphone, reflection and diffraction cause considerable changes in the pressure distribution about the microphone. Figure 7 shows the mean pressure over the membrane surface versus frequency for different angles of incidence. The response of the microphone

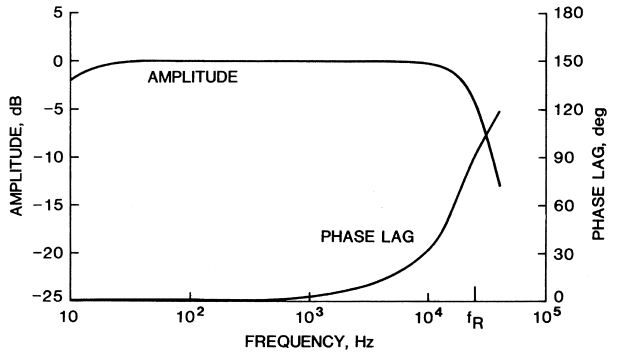


FIGURE 6 Pressure response of a $\frac{1}{2}$-in. air condenser microphone. At the resonant frequency f_R, the membrane displacement lags the incident pressure by 90°.

under this condition is called the free-field response. At normal incidence ($\phi = 0°$), the free-field effects are greatest; at grazing incidence ($\phi = 90°$), the pressure distribution is about the same as for the uniform pressure condition. The random incidence response curve can be regarded as the mean response when incidence from all directions is equally probable. A set of curves as shown in Fig. 7 can be used to correct the free-field response to an equivalent pressure response. A free-field microphone is intentionally made overdamped to yield the flattest frequency response for normal incidence. As a result, the upper cutoff frequency far exceeds that of the pressure response, sometimes by as much as a factor of 2. The upper cutoff frequencies shown in Table III are for pressure microphones.

7. Microphone Calibration

The three most widely used techniques for microphone calibration are the electrostatic actuator, the pistonphone, and the reciprocity procedure.

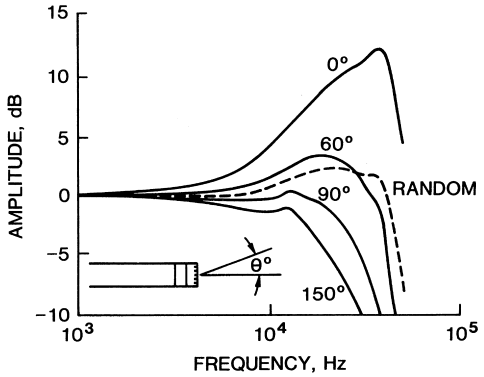


FIGURE 7 Free-field response of a $\frac{1}{2}$-in. air condenser microphone. (Redrawn with permission, courtesy of Bruel & Kjaer Instruments, Inc., Marlborough, MA.)

The electrostatic actuator is a flat metallic electrode positioned at a nominal distance d_1 from the microphone membrane. A voltage applied between the electrode and membrane produces a uniform electrostatic pressure on the membrane. If an ac voltage v_a is superimposed upon a high dc polarization voltage V_a, then the applied ac voltage and resulting electrostatic pressure have the same frequency. The membrane responds to the electrostatic pressure as it would to a sound pressure. The electrode is slotted to relieve acoustic loading between the electrode and membrane. The electrostatic pressure exciting the membrane is

$$p = \varepsilon_0 V_a v_a / d_1^2 \tag{14}$$

With typical values $V_a = 800$ V, $v_a = 30$ V rms, and $d_1 = 0.0005$ m, the rms pressure is 0.85 Pa, or 93 dB. The technique is excellent for obtaining frequency response but is not suitable for absolute calibration because of a twofold uncertainty in the distance d_1. First, the slots in the electrode necessitate a theoretically derived correction and, second, the polarization voltage shifts the equilibrium position of the membrane toward the actuator electrode.

The essential parts of a pistonphone are a coupler, into which the microphone cartridge is inserted and sealed, usually with an O-ring, and a vibrating piston of known displacement. The piston may be driven by a cam having a sinusoidal contour, generating a pure tone, or by a crankshaft, which in addition produces considerable second-harmonic distortion. The frequency is controlled through the speed of the cam or crankshaft. The pressure generated in the coupler is

$$p = \gamma P_0 S_p d / V \tag{15}$$

where S_p is the piston area, d the rms stroke, and V the volume of the coupler (including the equivalent volume of the microphone). With typical values of $\gamma = 1.4$, $P_0 = 10^5$ N/m^2 (1 atm), $S_p = 10^{-5}$ m^3, $d = 1.4 \times 10^{-4}$ m, and $V = 2 \times 10^{-5}$ m^3, the rms pressure is 9.8 Pa, or 114 dB. At audio frequencies, a precision of a couple of tenths of a decibel is attainable. At low frequencies a correction is needed for nonadiabatic compression. In the form of a portable, battery-operated device, it is ideally suited for quick calibration of microphones in the field. Two limitations are fixed amplitude and relatively low operating frequency (several hundred hertz maximum). Devices using moving-coil drivers (without servo control) are not true pistonphones, for the generated volume velocity depends on the acoustic impedance of the load.

The reciprocity procedure is based on the following electromechanical principle. When a reversible transducer is operated as a receiver, the ratio of open-circuit output voltage v to applied acoustic pressure p will equal some constant A. Then, when it is operated as a transmitter,

the ratio of the generated volume velocity U to the input current I will equal the same constant A, if the acoustic load Z_r is small (see Fig. 2a.a, b). According to this procedure, three transducers are placed pairwise in an acoustic coupler—a transmitter T, a reversible transducer R, and the test microphone M. In test 1, transmitter T generates a sound pressure P_T to excite receiver R, resulting in an open-circuit voltage $\upsilon_R = A_R P_T$. In test 2, test microphone M replaces transducer R to yield $\upsilon_M = A_M P_T$. These lead to the relationship:

$$A_M = A_R \upsilon_M / \upsilon_R \tag{16}$$

In test 3, transducer R, as transmitter, excites test microphone M, resulting in $\upsilon'_M = A_M P_R$. However, the above-stated reciprocity property, $U_R = A_R I_R$, and the known acoustic impedance of the coupler, $Z_C = P_R / U_R$, lead to the relationship:

$$A_R = \frac{U_R}{I_R} = \frac{P_R}{Z_C I_R} = \frac{\upsilon'_M}{A_M Z_C I_R} \tag{17}$$

Substitution of Eq. (17) into (16) yields:

$$A_M = (\upsilon_M \upsilon'_M / Z_C \upsilon_R I_R)^{1/2} \tag{18}$$

Thus, the microphone sensitivity depends only on electrical quantities, which are measurable to high precision, and a readily determinable acoustic impedance. The reciprocity method, the most precise of all known methods, can achieve absolute precisions of the order of hundredths of a decibel. A free-field variation of the procedure is similar but is beset with practical difficulties.

8. Microphone Performance in Harsh Environments

An increase in ambient temperature has three primary effects on condenser microphone parameters: a decrease in membrane tension, normally an increase in membrane–backplate gap, and an increase in air viscosity. The first two have compensating effects on the midband sensitivity. The last increases the membrane damping and is important only near resonance. As a result, the midband sensitivity of condenser microphones generally has a small temperature coefficient, typically <0.01 dB/°C over an interval from −50 to +60°C.

The air compliance C_A (Fig. 2) is inversely proportional and the air layer resistance R_A is directly proportional to the air density. Thus, a change in ambient pressure has little effect on midband sensitivity but has a strong influence on membrane damping. Typically, the pressure coefficient of the midband sensitivity is -10^{-5} dB/Pa.

Humidity is detrimental to condenser microphone performance primarily when it condenses and short-circuits the membrane to the backplate. Water vapor near saturation can cause arcing under an intense electric polarization field. A dehumidifier, containing a dessicant and inserted between the preamplifier and cartridge, has proved successful in keeping the interior of the cartridge dry. High relative humidities apparently do not affect the surface charge of an electret significantly. The immunity of the ceramic microphone from harmful effects of humidity is one reason for its popularity for many years.

The vibration sensitivity of a microphone depends on the direction in which the vibration is applied and is maximum when this direction is normal to the membrane surface. In this case, the equivalent acoustic pressure is

$$p = \rho_M t_M a_V \tag{19}$$

where a_V is the acceleration of the applied vibration. For a nickel membrane of density $\rho_M = 8850$ kg/m^3 and of thickness $t_M = 5 \times 10^{-6}$ m, an applied acceleration $a_V = 9.8$ m/sec (1 g) will produce an equivalent sound pressure of 0.43 Pa, or 87 dB. A low membrane surface density $\rho_M t_M$ is the key to suppressing vibration sensitivity.

In outdoor measurements, the ambient wind will generate considerable noise in a microphone and may disturb the intended measurement. A possible countermeasure is to install a windscreen, constructed of a fibrous material, either self-supporting or supported on a wire frame about the microphone. In principle, the material displays a high flow resistance to the quasi-static pressure of the wind but low resistance to acoustic pressures. A windscreen has two contrary effects, however. Its presence creates turbulence, an effect that can be minimized by making the dimensions of the windscreen sufficiently large, and it creates an acoustic cavity with excitable modes. The windscreen is most effective at high frequencies and in a wind direction normal to the membrane. Overall, in moderate winds (<30 km/hr) a windscreen may reduce the wind noise over the audio band by as much as 10–20 dB.

Often it is necessary to make sound pressure measurements in extremely hostile or inaccessible locations, for example, in jet engine exhausts or in the ear canal. Such measurements can be realized with the aid of a probe tube—a long, thin, hard-walled tube of the general configuration shown in Fig. 8. Ideally, the sound pressure p_m at the microphone, coupled to one end of the tube in a coupler cavity, will be the same as the test pressure p_T at the probe tip. This will be approximately the case at long acoustic wavelengths or when the load impedance at one end matches the characteristic impedance of the tube. Otherwise, the natural tube resonances will cause an undulating frequency response. Since the characteristic tube impedance is resistive, impedance matching can be achieved by means of an acoustic damping material placed at the probe tip. The response of the probe tube for the underdamped, correctly damped, and overdamped cases is shown at the bottom of the figure.

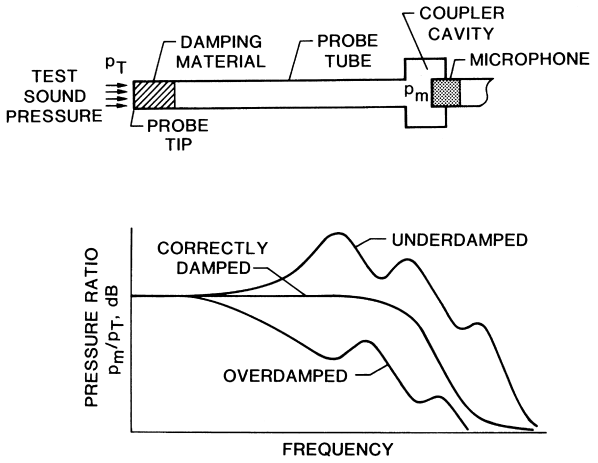


FIGURE 8 Probe tube and its response when it is under damped, correctly damped, and overdamped. (Redrawn with permission, courtesy of Bruel & Kjaer Instruments, Inc., Marlborough, MA.)

9. Infrasonic Measurements

There are two reasons for the low-frequency rolloff of a condenser microphone using a polarization voltage: one electrical and the other mechanical. The electrical rolloff is due to the charging time constant of the circuit of Fig. 3, for which the output voltage is

$$v = \frac{\Delta C V_0}{C_E} \frac{j\omega R_E C_E}{1 + j\omega R_E C_E} \tag{20}$$

where $C_E = C + C_s + C_i$ and $R_E = R \| R_C$. The electrical cutoff frequency is $(2\pi R_E C_E)^{-1}$.

The origin of the mechanical rolloff lies in the capillary vent tube shown in Fig. 1. The vent is necessary for static pressure equalization on both sides of the membrane; otherwise, the microphone will show an undesirable response to changes in ambient pressure. The mechanical cutoff frequency is $(2\pi R_V C_V)^{-1}$, where R_V is the acoustic resistance of the capillary tube and C_V the acoustic compliance of the backchamber. Both the electrical and mechanical cutoff frequencies can be controlled through the choice of design components and are usually set equal to one another.

For infrasonic measurements, the mechanical rolloff can be eliminated by closing the vent tube and the electrical rolloff through the use of a microphone carrier system. Here, the microphone is made the capacitive element in a tank circuit, which is tuned to an electrical carrier frequency, typically 1–10 MHz. Motion of the membrane detunes the tank circuit—an effect that is used to amplitude- or frequency-modulate the carrier voltage. Such a system responds to static changes in microphone capacitance, in other words, to frequencies down to dc. A voltage-controlled capacitor placed across the microphone capacitance allows provision for automatic feedback compensation of capacitance changes due to changes in ambient pressure, as well as for remote calibration of the electronic system (the "insertion" technique). However, the time constant of the feedback control system places a lower limit on the measurable signal frequencies. A carrier system has a relatively high noise floor, typically 50 dB over 20 Hz to 20 kHz for a 1-in. microphone, but a wide frequency response, from dc to about 20% of the carrier frequency, microphone permitting. An infrasonic microphone system is best calibrated by an infrasonic pistonphone, but the coupler must be extremely leak tight.

10. Fiberoptic Sensors

The transmission of acoustically generated signals through optical fibers has two major advantages over their copper counterparts. The first is immunity from electromagnetic interference, thus dramatically reducing the practical problems associated with grounding, shielding, and guarding. The second is the remote placement of the supporting optoelectronics, permitting the sensing element (e.g., membrane) to operate in harsh environments and confined locations. Classification of fiberoptic sensors follows the property of light that is modulated: wavelength (or phase), intensity, or polarization. Phase-modulating sensors are further classified into grating and interferometric sensors, and intensity-modulating sensors are classified according to whether the modulation affects the guided or evanescent light wave. Polarization-modulating sensors have not enjoyed the comparable level of development of the others and will not be discussed further. As a rule, interferometric sensors employ single-mode fibers and itensity-modulating sensors multimode fibers. Sensors can be designed to serve as microphones or hydrophones.

In the *Mach-Zehnder interferometer* (Fig. 9a) light from the source splits at the first coupler, one beam passing through the sensor fiber and the other through the reference fiber. They recombine at the second coupler and are detected at the photodetector. A sound wave incident upon the sensor fiber modulates the phase of the sensor beam relative to that of the reference beam. The resulting temporal interference produces an optical signal proportional to the acoustical excitation.

The *Fabry-Perot interferometer* (Fig. 9b) passes the source light through a coupler, where a fraction, say 50%, continues toward the membrane, the remainder lost at the absorption cell (to prevent reflection back to the coupler). The membrane and the end of the first fiber comprise a Fabry-Perot cavity, thus generating an interference pattern which is modulated by the sound wave incident upon the membrane. The modulated light returns to the coupler and 50% passes on to the photodetector.

In the intensity-modulating *fiberoptic lever* (Fig. 9c), light passing through a bundle of transmitting fibers is

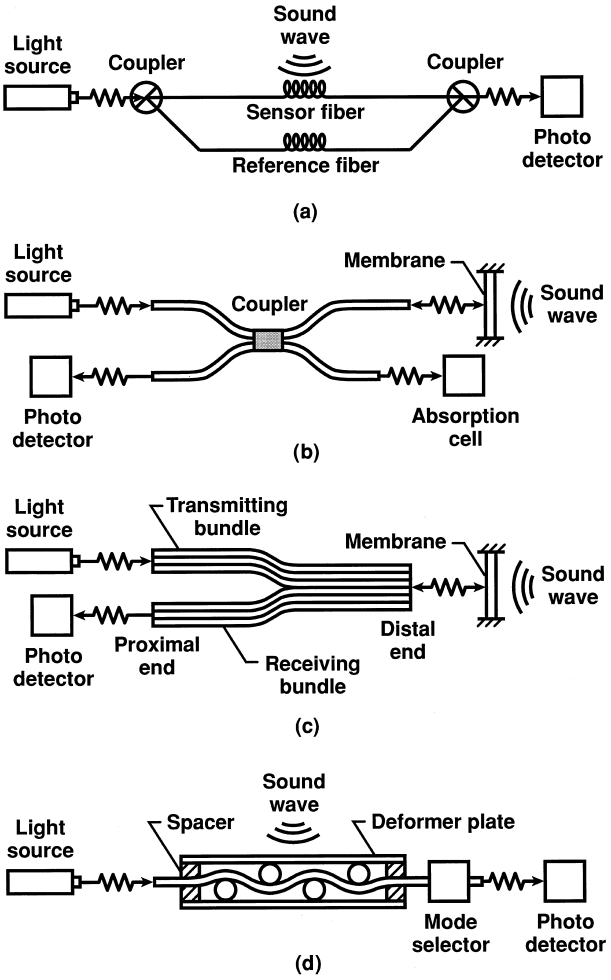


FIGURE 9 Fiberoptic sensors. (a) Mach-Zehnder interferometer; (b) Fabry-Perot interferometer; (c) fiberoptic lever; and (d) microbend sensor.

reflected from the membrane. A fraction of the reflected light is intercepted by a bundle of receiving fibers. A sound wave incident upon the membrane modulates the fraction of received light, thus the intensity of light into the receiving bundle and photodetector.

In the *microbend* sensor (Fig. 9d), the optical fiber is constrained to a periodic deformation. The periodicity determines the coupling between the optical propagating modes and radiating modes (exiting the fiber). An incident sound wave upon the deformer plate modulates the fiber deformation, the modal coupling, and consequently the light to the photodetector.

In general, the sensitivity of a fiberoptic pressure sensor is the product of three component sensitivities: (1) mechanical, change of sensing element displacement per unit sound pressure; (2) optical, change of optical signal per unit sensing element displacement; and (3) electronic, change of output voltage per unit optical signal in the photodetector. In a well-designed sensor, the threshold sensitivity is limited by the shot noise in the photodetector.

B. Measurement of Sound Intensity and Power

1. Sound Intensity

Sound intensity is the sound energy passing through a unit normal area per unit time. It is a vector quantity having units of watts per square meter (W/m^2). This definition can be expressed in terms of the following fundamental acoustic parameters,

$$\bar{I} = \langle p(t)\bar{u}(t)\rangle \tag{21}$$

where the pressure and particle velocity are time-varying quantities and $\langle\,\rangle$ denotes a time average. For harmonic waves:

$$\bar{I} = \mathrm{Re}\{\mathbf{p}\bar{\mathbf{u}}^*\} = \mathbf{p}\bar{\mathbf{u}}\cos\theta \tag{22}$$

where boldface symbols denote rms time averages; Re, the real part; the asterisk, a complex conjugate; and θ, the temporal phase angle between $\mathbf{p}$ and $\bar{\mathbf{u}}$. For illustration, let us apply Eq. (22) to two simple examples. For a plane wave, $\mathbf{u} = \mathbf{p}/\rho c$, $\theta = 0$, and $I = \mathbf{p}^2/\rho c$, where ρ is the air density and c the sound velocity. For a standing wave, $\mathbf{u} = \mathbf{p}/\rho c$, $\theta = \pi/2$, and $I = 0$. To avoid the difficult measurement of sound particle velocity, we invoke Newton's second law:

$$u = -\rho^{-1}\int \nabla p\, dt \tag{23}$$

where ∇p is the pressure gradient. Substituting Eq. (23) into (21) yields an expression in p alone:

$$\bar{I} = -\left\langle (p/\rho)\int \nabla p\, dt\right\rangle \tag{24}$$

In a practical measurement system, two microphones are aligned coaxially face to face, or in a coplanar arrangement, and separated by a fixed distance Δr in the sound field (Fig. 10). As long as $\Delta r \ll \lambda$, the wavelength, then Eq. (24) can be approximated as follows:

$$I_r = -\frac{1}{\rho}\left\langle \frac{(p_A + p_B)}{2}\int \frac{(p_A - p_B)}{\Delta r}\, dt\right\rangle \tag{25}$$

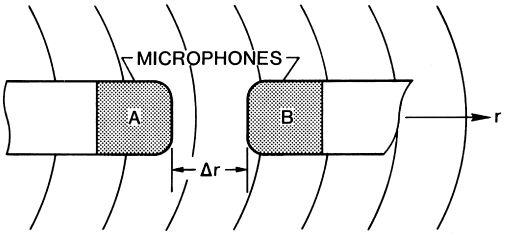


FIGURE 10 Measurement of sound intensity with a two-microphone arrangement. The microphones can also be aligned with coplanar membranes.

Note that p is represented by the mean value of p_A and p_B. Following the measurement of the sound pressures p_A and p_B, the sound intensity is evaluated by the signal-processing operations of summing, integrating, multiplying, and time averaging. It is good practice to repeat a measurement with the microphones switched in order to cancel the effect of phase differences between channels. Equation (25) yields the intensity component along the microphone axis. For sound propagation in a direction α relative to this axis, I_r must be divided by $\cos\alpha$.

A second measurement method is based on the fact that the time-averaging operation indicated in Eq. (24) is closely related to the crosspower spectrum G_{AB} between p_A and p_B. Formal analysis yields:

$$I_r = -(\omega\rho\Delta r)^{-1}\,\mathrm{Im}\{G_{AB}\} \tag{26}$$

where Im denotes the imaginary part. If the microphone signals are applied to a correlator or digitally to a computer, Eq. (26) can be used to compute the intensity.

Applications of sound intensity include sound power measurement (see Section I.B.3) and intensity mapping to locate sound sources and sinks.

2. Acoustic Enclosures

Before discussing sound power measurement, we shall find it helpful to discuss four basic types of acoustic enclosure: anechoic, reverberant, resonant, and Helmholtz. These are depicted in Figs. 11a–d.

An anechoic chamber has no reflections (echoes) from its walls and thus simulates free-field conditions. The anechoic condition is realized by lining the walls of the chamber with wedges made of a porous, absorptive material such as rock wool, glass wool, or foam. Typical wedge dimensions are 20–60 cm at the base and a wedge angle of 10–15°. Since the floor must also be lined, an enclosure usually has a wire mesh just above the wedge tips to support personnel and equipment. The free-field approximation is truest at high frequencies. The lowfrequency limit is determined by the criterion:

$$h/\lambda > (1 + R_w)/(1 - R_w) \tag{27}$$

where h is the appropriate chamber dimension and R_w the reflection coefficient of the wedge along its axis. Using typical values $R_w = 0.1$ and $h = 10$ m, we find $\lambda < 8.33$ m, corresponding to a minimum operating frequency of 40 Hz. The background noise level of a good anechoic chamber may lie lower than 0 dB SPL over most of the entire audio range. However, the mark of quality is how well the sound pressure from a point source adheres to the $1/r$ spherical spreading law, which is perturbed by reflections from the walls. Within confines no closer than about a meter from the walls, verification of the law to within 1 dB is a reasonable design goal.

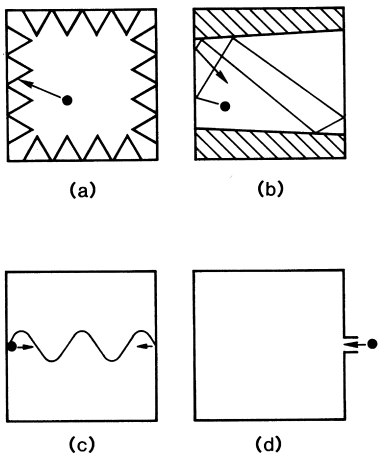


FIGURE 11 Acoustic enclosures. (a) Anechoic: Waves from the sound source (dark circle) are absorbed without reflection. (b) Reverberant: Waves experience multiple reflections to establish a uniform sound energy density. (c) Resonant: Reflected waves reinforce waves generated from the source to establish a standing wave. (d) Helmholtz: At long wavelengths, the chamber behaves as a compliance and its small opening as an acoustic mass.

A reverberation chamber is used to produce ideally a spatially uniform sound energy density. The walls are hard and highly reflective, such that a sound ray emitted from a source will experience multiple reflections in haphazard fashion, as shown in Fig. 11b, and will eventually fill the room. When the source is turned on, the energy density builds up until dissipation balances the sound power emitted by the source. The resulting sound field is called a diffuse field, independent of location or direction, a situation difficult to achieve in practice. Hard reflecting objects and moving reflectors enhance the diffuseness. The volume V of the room should exceed $3\lambda^3$ for octave analysis and $9\lambda^3$ for third-octave analysis (see Section II.A.2), where λ is the largest wavelength. Recommended ratios for room dimensions are $1 : 2^{1/3} : 4^{1/3}$. The source should be located at least $\frac{1}{4}\lambda$ from the walls, and microphones at locations removed from known peaks and valleys in sound pressure. The reverberation chamber is widely used for measurement of sound absorption of materials and sound power emission of sources.

An enclosure is said to be resonant if a reflected wave returns to the source, in the direction from which it was emitted, after progressing an integral number of wavelengths. The returning waves reinforce the emitted waves to produce a standing wave pattern, which grows in amplitude until dissipation equals emission. In Fig. 11c, a plane wave is reflected between parallel walls. The resonant frequency for this one-dimensional case is

$$f_R = c/\lambda = nc/2h \tag{28}$$

where h is the distance between the walls and n an integer. In two and three dimensions, the standing wave patterns can become quite complex. Resonance conditions also exist for bound cylindrical and spherical waves.

A Helmholtz resonator is the acoustic analog of a mass–spring system. Two acoustic elements are coupled together: a chamber and an opening in the form of an orifice or tube. At wavelengths much greater than the element dimensions, the air in the opening moves as a unit and thus behaves as an acoustic mass; the chamber behaves as a compliance. The resonant frequency of the mass–compliance system is

$$f_R = cS/lV \tag{29}$$

where S is the cross-sectional area and l the effective length of the opening, and V is the volume of the chamber. This frequency is much lower than the natural frequencies of the chamber alone. Helmholtz resonators are very common in nature and technology, from caves to wine jugs to leaky rooms, and at one time were used to determine pitch. We have already come across an example: the backchamber–vent hole system of a condenser microphone.

3. Sound Power

Sound power is the integrated normal intensity over a surface enclosing an acoustic source and has units of watts:

$$W = \iint_S \vec{I} \cdot d\vec{\mathbf{S}} \tag{30}$$

In the absence of absorption the sound power is independent of distance from the source. Three methods for measuring sound power are free field, diffuse field, and reference source.

The free-field method assumes that the sound field is (1) far field, (2) spherically spreading, and (3) in an anechoic environment. The sound source must be located either outdoors or in an anechoic chamber in order to approximate free–field conditions. Microphone measurement stations are located on an imaginary sphere (source-suspended) or hemisphere (source-grounded), having a radius r several times greater than the longest wavelength. The sound power is determined from the mean value p_m^2 of the squares of the sound pressure measurements at all the microphone stations:

$$W = \left(p_m^2/\rho c\right)(4\pi r^2)F \tag{31}$$

The directivity factor F has the value 1 if the source is suspended or 0.5 if the source is grounded on a perfectly reflecting surface. At 20°C, 1 atm, $\rho c = 415$ N sec/m^5 in air. For example, if $p_m = 20\,\mu$Pa, $4\pi r^2 = 1$ m^2, and $F = 1$, then $W = 10^{-12}$ W.

If the sound intensity, using microphone pairs, is measured instead of the sound pressure alone, the requirements on far-field, spherically spreading sound propagation can be relaxed. However, so far this modification has not gained widespread usage.

The diffuse-field method is based on the fact that the sound energy density is uniform throughout a reverberant enclosure. If the sound source is placed in a good reverberation chamber, then the energy density adjusts to a steady-state condition, whereby the sound power emitted by the source balances the power dissipated in the chamber. The experimental arrangement is similar to that used in free-field measurements, but now the sound power becomes:

$$W = p_m^2 R/4\rho c \tag{32}$$

where R is the reverberation constant of the room and can be determined from the measurement of reverberation time (see Section III.A).

Often the enclosure about a sound source is neither anechoic nor reverberant, and accommodation to acoustic requirements is not possible—for example, if the source is a heavy machine. In such a case, the reference source method may prove useful. Here, an identical set of measurements is taken both for the test source and for a reference source of known sound power. Then,

$$W = W_r p_m^2/p_{mr}^2 \tag{33}$$

where W_r and p_{mr}^2 are the sound power and mean-squared sound pressure for the reference source.

C. Measurement of Acoustic Particle Velocity

This difficult measurement is generally avoided if the acoustic particle velocity can be related simply to the sound pressure. Otherwise, one approach is to exploit the relationship between particle velocity and sound pressure gradient given in Eq. (23).

1. Pressure Gradient Microphone

Measurement of a sound pressure gradient can be achieved by any of the principles used to measure sound pressure alone. A highly successful device is the foil–electret gradient microphone of Sessler and West, illustrated in Fig. 12. It is assumed that the microphone is sufficiently small not to perturb the sound field, requiring $\Delta l < \lambda/2$, and that the sound pressures on either side of the membrane are the same as those at the protective grids, namely, p_1 and p_2.

A metallized electret foil is stretched tightly across, and in contact with, the backplate. Thus the response is compliance-controlled, although a mass-controlled design is feasible. An air gap exists only because of irregularities in the backplate surface. Holes in the backplate and both protective grids permit the sound wave to gain access to

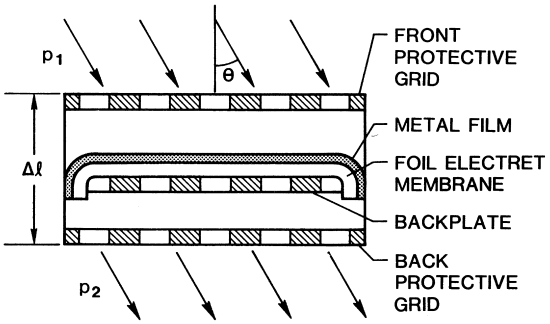


FIGURE 12 Foil–electret gradient microphone. This device can be used to measure acoustic particle velocity.

both sides of the membrane. The membrane responds to the net normal component of sound pressure,

$$p_i = |-\nabla p \Delta l' \cos\theta|$$

which is related to sound particle velocity u through Eq. (23). For a harmonic wave of frequency f:

$$p_i = 2\pi f u \rho_0 \Delta l' \cos\theta \tag{34}$$

where ρ_0 is the static density of air and $\Delta l \approx \Delta l'$ the effective grid spacing. Measurement of p_i thus determines u since all other quantities in Eq. (34) are known. Measurement of the response at constant amplitude but increasing frequency is a good check as to whether the microphone is truly responding to pressure gradient.

2. Other Methods

The Rayleigh disk is based on the fact that a sound wave will exert a torque on a suspended disk oriented at a suitable angle, usually 45°, with respect to the sound field. The torque is proportional to the square of the incident sound particle velocity. In a practical apparatus the suspending thread exercises a resisting torque, which balances the sound-generated torque. Measurement of the angular displacement, usually by optical means, permits determination of the torque and thus particle velocity. Such measurements are most conveniently conducted in a tube since the direction of sound propagation is known. At one time, the British Post Office used this method for the absolute calibration of microphones.

The hot-wire anemometer is not suitable for purely acoustic excitation but has been used successfully for acoustic excitation superimposed on a mean flow.

D. Measurement of Sound Speed and Attenuation

1. Sound Speed

The distance per unit time through which a phase point of a sound signal propagates is the sound speed and, with specified direction, is called the phase velocity. It is a physical property of the medium, although it has a slight dependence on the frequency. The theoretical expression for the sound speed in a gas at low pressures, $c = (\gamma RT/M)^{1/2}$, where γ is the ratio of specific heats at constant pressure and constant volume, R the universal gas constant, T the absolute temperature, and M the molecular weight, is very reliable. Measurements of sound speed are used to determine the (nonideal) equation of state and the second virial coefficient. Measurements are usually taken at low frequencies (low kilohertz range) because thermoviscous (classical) absorption increases very strongly with frequency. In a liquid $c = (\kappa\rho)^{-1/2}$, where κ is the adiabatic compressibility and ρ the density, classical absorption is not so great a problem and sound speed shows minimal dispersion well into the high megahertz region. Measurements are used to determine the compressibility and the nonlinearity parameters. Of course, the sound speed is also of interest to acousticians engaged in sound propagation problems. At 20°C, 1 atm, the speed of sound is 343.23 m/sec in dry air, 1482.34 m/sec in pure water, and 1529.03 m/sec in seawater (3.5% salinity). Two apparatuses commonly used for the measurement are the cylindrical resonator and the spherical resonator.

2. Cylindrical Resonator

This resonator is a long hollow cylinder capped by two parallel end plates. It employs a transmitter and receiver, either separately or as a single unit and usually located in the end plates. Many types of electroacoustic transducers are used: piezoelectric, electrostatic, electrodynamic, and so on. If the excitation is in the form of a short pulse or tone burst (see Section I.D.4), then $c = L/T$, where L is the distance from source to receiver and T the corresponding transit time. If an axial mode of the resonator is excited, then:

$$c = 2Lf_n/n, \qquad n = 1, 2, \ldots \tag{35}$$

The advantages of the cylindrical geometry are modal purity and ease of fabrication, especially for measurements at high pressures. The disadvantages are the relatively large thermoviscous losses at the walls, which require a correction to the sound speed, and the existence of cutoff frequencies, which limit the maximum usable frequency. The first nonaxial mode has a diametric node at a frequency:

$$f = 0.5861c/2R \tag{36}$$

where R is the internal radius. If the resonator is symmetrical about the cylindrical axis, this mode will not be strongly excited and the limiting frequency will be that

corresponding to the first radial mode:

$$f = 1.2187c/2R \tag{37}$$

For example, if $L = 1$ m, $R = 0.025$ m, and $f_1 = 600$ Hz (measured), then Eq. (35) yields $c = 300$ m/sec. Ignoring the mode defined by Eq. (36), we find the cutoff frequency from Eq. (37) to be $f = 7312$ Hz, which permits the use of $7312/600 \approx 12$ axial modes.

3. Spherical Resonator

The excitation of a radial wave in a spherical resonator has an important advantage: The particle velocity is normal to the walls, and the wall losses are typically 10 : 1 lower than those of a cylindrical resonator at a comparable frequency. The major difficulty lies in fabricating a spherical shell with precise internal dimensions. Small glass spheres of volumes from 0.5–100 liters permit measurement in liquids from about 5 kHz to 2 MHz. Larger resonators invariably consist of two metallic hemispheres joined by flanges at the equator. The sound speed is related to the radial mode frequencies by:

$$c = 2\pi R f_n / \nu_n \tag{38}$$

where R is the radius of the sphere and $\nu_n = 4.493, 7.725, 10.904, \ldots$ for $n = 1, 2, 3, \ldots$. Geometric imperfections affect c only to second order. Measurement precisions of 0.02% are commonplace; precisions of 0.003% have been achieved.

4. Sound Attenuation

Sound attenuation is the reduction in amplitude of a sound wave as it propagates through a medium. It may be the result of spreading, scattering, or absorption (direct conversion to heat). The same apparatus can be used to measure both speed and attenuation; often both quantities are measured together.

There are many ways to designate sound attenuation, most associated with a particular experimental technique. The quality factor Q is the most fundamental measure because it is based on energy considerations. If a medium is subjected to periodic acoustic excitation, then:

$$Q = \frac{2\pi \times \text{maximum energy stored}}{\text{energy dissipation per cycle}} \tag{39}$$

The quality factor—a physical property of the medium, just as the density and compliance are—is sensitive to physical and chemical changes, sometimes strongly frequency dependent, but ill-suited to direct measurement. Rather, Q is determined through its relationship to the other measures of attenuation shown in the following tabulation.

Attenuation constant, Np/m	$\alpha = \pi/\lambda Q$
Reciprocal time constant, Np/sec	$\beta = \pi f/Q$
Logarithmic decrement, Np/wavelength	$\Lambda = \pi/Q$
Resonant halfwidth, Hz	$\Delta f = f_0/Q$
Tangent of phase angle	$\tan\delta = 1/Q$

When the wavelength is much smaller than the dimensions of the experimental enclosure—and this includes propagation in the free field—the attenuation constant α is the appropriate measure. A popular method of determining α is the pulse–echo method, illustrated in Fig. 13a. A transmitter T launches a tone burst, a packet of harmonic waves of frequency f, in a cylindrical resonator. In the sequence of reflections from the end plates, a receiver R at a fixed station measures the amplitude of the packet, which attenuates by a factor $\exp(-\alpha x)$ over a propagation distance x. A suitable pulse width must be chosen by the observer: Too great a width will decrease spatial resolution, but too small a width will decrease frequency resolution. More sophisticated techniques fall into the domain of ultrasonic measurements. The attenuation constant α is the most common measure of attenuation in fluids and the one normally shown in tabulations of absorption values.

The attenuation can also be described in terms of a time constant (or its reciprocal), as is common in reverberation measurements. In a typical experiment, white noise is used

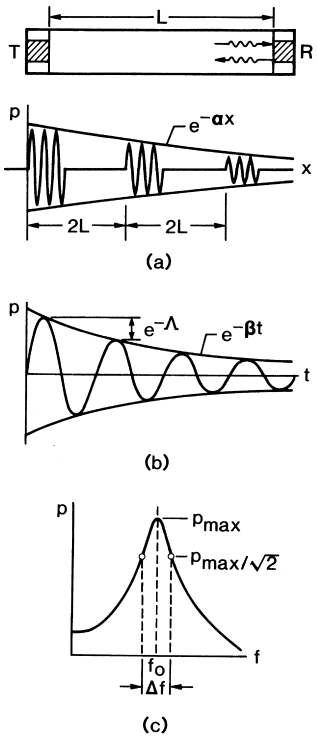


FIGURE 13 Measurement of sound attenuation by (a) the pulse–echo, (b) the free decay, and (c) the resonant halfwidth methods.

to excite a large number of reverberation chamber modes, which can be observed individually on a fixed receiver through narrow band filtering. When the excitation is removed, the amplitude decay, of the form $\exp(-\beta t)$, is observed on each filter. The decay envelope is shown in Fig. 13b.

The next three measures pertain to the resonant technique, whereby a natural mode of a resonator is excited, usually an axial mode of a cylinder or radial mode of a sphere.

When the excitation is removed, the sound pressure amplitude of the mode decays freely, as in the reverberation technique. The logarithmic decrement Λ is the natural logarithm of the amplitude ratio of two successive peaks, shown in Fig. 13b:

$$\Lambda = \ln(p_n/p_{n+1}) \tag{40}$$

Usually Λ is averaged over a range of the free decay, typically 10–40 dB. This method is most suitable for media with a high Q, say $Q > 10$: in media with an extremely high Q, such as degassed distilled water for which $Q > 10^6$, this is the only feasible method for measuring sound attenuation.

In a steady-state experiment, the attenuation can be determined by the halfwidth of the resonance curve, shown in Fig. 13c. The sound pressure amplitude is measured as the frequency is incremented or swept through the resonant frequency f_0. The halfwidth Δf is defined as the frequency interval between the two points where $p = p_{\max}/\sqrt{2}$; that is, where the amplitude is 3 dB down from the peak $p_{\max}$. This method cannot be used when the halfwidth is too sharp, due to instability of f_0, or when the halfwidth is too broad, due to poor peak definition. Furthermore, any losses inherent in the transmitter itself will contribute to the halfwidth. With an efficient transmitter this method can be used effectively over a range of Q from about two to several hundred.

If the medium is very lossy ($Q < 2$), the most effective measure of attenuation is the loss tangent $\tan\delta$. Such low Q's are found in some polymeric liquids, seldom in gases. Because of the high loss the sample is made the lossy element of a composite resonator, in which independent measurements of force and displacement yield the phase angle δ.

The measured attenuation in a resonator contains three principal components: wall absorption, absorption due to fluid–structure interaction, and the constituent absorption that is to be measured. The wall and structural components, called the background absorption, can be determined through measurements on a background fluid having negligible constituent absorption over the range of measurement parameters. In gases, argon and nitrogen are frequently used for this purpose.

5. Free-Field Measurements of Attenuation

In the free field, corrections must be made for spreading. For a spherical source, the sound pressure falls as $1/r$. At sufficiently large distances from the source, a spherical wave can be approximated by a plane wave, and the correction is not needed.

6. Optoacoustical Method: Laser-Induced Thermal Acoustics

The passage of laser light through a fluid can induce a strain either thermally (resonant) or electrostrictively (nonresonant). A typical laser-induced thermal acoustics (LITA) arrangement is shown in Fig. 14. Typical component specifications are shown in parentheses. Light from a pulsed pump laser ($\lambda_{\text{pump}} = 532$ nm) is split into two beams which intersect at a small angle ($2\theta = 0.9^\circ$). Optical interference fringes of spatial period $\Lambda = \lambda_{\text{pump}}/(2\sin\theta)$ generate electrostrictively counterpropagating ultrasonic waves of fixed wavelength to form a Bragg grating, shown in the insert. A long-pulsed probe laser (750 nm) illuminates the grating, which diffracts a small fraction of the probe beam at an angle ϕ to a photomultiplier. The diffracted signal is normalized to the direct probe signal measured at the photodetector. Since the acoustical wavelength is known from the intersection angle 2θ and

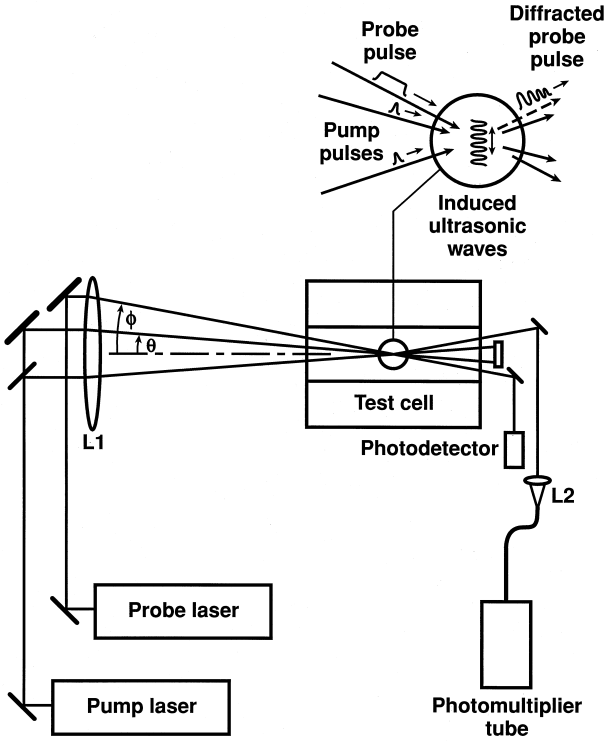


FIGURE 14 Laser-induced thermal acoustics.

pump laser wavelength, and the frequency is known from the photomultiplier signal, the speed of sound of the fluid medium can be measured. A "referenced" version of LITA, implemented to avoid the large error associated with the intersection angle measurement, splits the pump and probe beams and directs them to a second LITA cell containing a fluid of known sound speed.

E. Measurement of Acoustic Impedance

The relationship between sound pressure and acoustic flow velocity plays a central role in the analysis of acoustic devices, such as mufflers and musical instruments, and in the determination of a sound field in the presence of a boundary. Quantitatively, this relationship is described by one of three types of acoustic impedance. The acoustic impedance—the ratio of sound pressure to volume velocity $Z = p/U$—is used in analyzing acoustic circuits, where devices are represented by equivalent lumped elements. It is a property of the medium, frequency, and geometry and has units of $\mathrm{N \cdot sec/m^5} = \mathrm{kg/sec \cdot m^4} = \mathrm{Rayl/m^2}$.

The interaction between a sound wave and a boundary depends on the specific acoustic impedance of the boundary relative to that of the propagation medium. The specific acoustic impedance is the ratio of sound pressure to acoustic particle velocity $z = p/u$. For a plane wave $z = \rho c$, and it is basically a property of the medium, although it can have complex frequency-dependent parts. It is related to Z through $z = ZS$, where S is the cross-sectional area and has units of $\mathrm{N \cdot sec/m^3} = \mathrm{kg/sec \cdot m^2} = \mathrm{Rayl}$. Thus measurement of z readily leads to the determination of Z and vice versa.

The mechanical or radiation impedance, the ratio of force to particle velocity, is of interest in systems containing both discrete and continuous components but is not discussed here.

Methods of measuring acoustic impedance fall into three broad categories: (1) impedance tube and waveguide methods in general, (2) free-field methods, and (3) direct measurement of sound pressure and volume velocity.

1. Impedance Tube

The test specimen is located at one end of a rigid tube and a transmitter at the other end (Fig. 15a). It is important to distinguish between materials of local reaction and those of extended reaction. In the former, the behavior of one point on the surface depends only on excitation at that point and not on events taking place elsewhere in the material. In the latter, acoustic excitation at a point on the surface generates waves that propagate laterally throughout the material. Generally, a material is locally reacting if normal acoustic penetration does not exceed a wavelength.

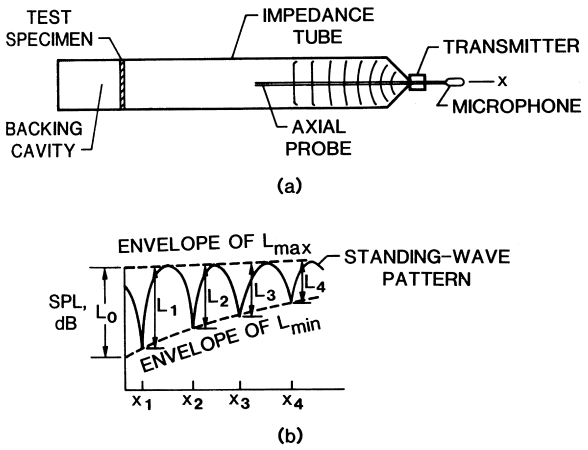


FIGURE 15 Measurement of acoustic impedance with an impedance tube. (a) Impedance tube, and (b) standing wave pattern and its envelopes.

For a locally reacting material, a thin test specimen (thickness $\ll \lambda/4$) is backed by a $\lambda/4$ air gap sandwiched between the specimen and a massive reflector. For a material of extended reaction, a specimen of approximate thickness $\lambda/4$ is backed by the massive reflector directly against its surface.

The transmitter is tuned to establish a standing wave pattern, which is probed by a microphone located either within the tube or at the end of a probe tube. The observer slides the probe along the impedance tube axis and records the standing wave pattern $L(x)$ in decibels (Fig. 15b). Here $L(x) = 20 \log[p(x)/(p_0)]$, where the reference pressure p_0 is immaterial. The impedance is evaluated from the pressure standing wave ratio L_0 at the specimen surface, which cannot be measured directly but is computed from a best fit to the trend of $L_{\max}$ and $L_{\min}$ shown in the figure. After computing the antilog of the pressure standing wave ratio K_0 and related quantities,

$$K_0 = 10^{L_0/20} \tag{41}$$

$$\phi = \left(\frac{x_1}{x_2 - x_1} - \frac{1}{2}\right) \times 360^\circ \tag{42}$$

$$M = \tfrac{1}{2}\left(K_0 + K_0^{-1}\right) \tag{43}$$

$$N = \tfrac{1}{2}\left(K_0 - K_0^{-1}\right) \tag{44}$$

we determine the real z' and imaginary z'' parts of the specific acoustic impedance relative to that of air, ρc:

$$\frac{z'}{\rho c} = \frac{1}{M - N\cos\phi} \tag{45}$$

$$\frac{z''}{\rho c} = \frac{N\sin\phi}{M - N\cos\phi} \tag{46}$$

This method is capable of yielding measurements of high precision, to within a few percent based on repeatability.

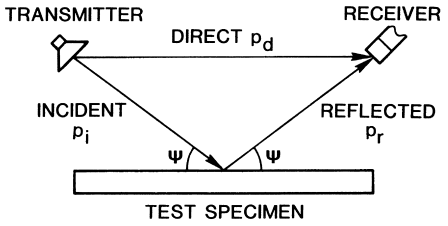


FIGURE 16 Measurement of acoustic impedance by a free-field method.

A major source of error lies in the determination of x, which may be illdefined for a rough or fibrous specimen surface. To improve surface definition, a face sheet composed of a fine-meshed gauze of low acoustic resistance can be used. A disadvantage is the time required to take the number of measurements needed to establish the standing wave pattern. More modern methods based on the transfer function between two microphone stations reduces the measurement time considerably.

2. Free-Field Methods

A transmitter sends an incident wave at an angle ψ toward the test specimen, from which it is reflected, also at an angle ψ, toward a receiver (Fig. 16). The specific acoustic impedance is evaluated from the reflection coefficient $R_p = p_r/p_i$:

$$\frac{z}{\rho c} = \frac{1}{\sin\psi}\left(\frac{1+R_p}{1-R_p}\right) \tag{47}$$

A variety of techniques, both transient and steady state, have been devised to determine the three wave components p_r, p_i, and p_d. One steady-state method utilizes three separate measurements at each frequency: (1) with the specimen in place, yielding $p_1 = p_r + p_d$; (2) with the specimen replaced by a reflector of high impedance, yielding $p_2 = p_r' + p_d \approx p_i + p_d$; and (3) with the reflector removed, yielding $p_3 = p_d$ alone. Thus,

$$R_p = \frac{p_r}{p_i} = \frac{p_1 - p_3}{p_2 - p_3} \tag{48}$$

Free-field methods are used for testing materials at short wavelengths and are popular for outdoor measurements of the earth's ground surface.

3. Direct Measurement of Sound Pressure and Volume Velocity

For measurement of the acoustic impedance within an acoustic device, the sound pressure can be measured with the aid of a probe tube (Section I.A.8), but measurement of the acoustic particle or volume velocity is difficult (Section I.C).

The most common method of attacking the latter problem is to control the volume velocity at the transmitter. This can be achieved in several ways: (1) by mounting a displacement sensor on the driver; (2) by using a dual driver, directing one side to the test region and the other side to a known impedance Z_k and using $U = p/Z_k$ and (3) by exciting a driving piston with a cam so that the generated volume velocity will be independent of acoustic load. The first two methods rely on the integrity of the velocity measurement technique: the third is limited to relatively low frequencies. To measure the specific acoustic impedance of a material, a transmitter, receiver, and test specimen are mounted in a coupler; the impedance of the latter must be taken into account.

II. INSTRUMENTS FOR PROCESSING ACOUSTICAL DATA

A. Filters

The representation of an acoustic time history in the domain of an integral (or discrete) transform has two advantages. First, it transforms an integrodifferential (or difference) equation into a more tractable algebraic equation. Second, it often separates relevant signal from irrelevant signal and random noise. The two most common transforms used in acoustics are the Fourier transform for continuous time histories and the z transform for discrete (sampled) time histories. The Fourier transform represents a time history $f(t)$ in the frequency domain,

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\exp(-j\omega t)\,dt \tag{49}$$

with $\omega = 2\pi f$. The z transform represents the sampled values $f(nT_s)$ in the z domain:

$$F(z) = \sum_{n=0}^{\infty} f(nT_s)z^{-n} \tag{50}$$

where T_s is the sample interval and n the sample number. The representation of a time history in the transformed domain is called a spectrum. We shall be concerned with the frequency spectrum.

Filters fulfill three major functions in acoustics: spectral selection, analysis, and shaping. It is assumed that the reader is familiar with the general characteristics of filters and with filter terminology.

1. Spectral Selection

We shall present two examples of spectral selection. The first is antialiasing. In sampled systems, it is essential that all frequency components above half the sampling frequency f_s be suppressed to avoid "aliasing," that is, the appearance of components of frequency $f_s - f$ in the observed spectrum. This is a consequence of the Nyquist sampling theorem. The maximum frequency for which

the spectrum is uncorrupted by aliaising is called the Nyquist frequency. The second example is signal-to-noise improvement. The observed signal is often a pure tone, for which narrow-band filtering will produce a considerable improvement in signal-to-noise (S/N) ratio. If the noise is "white," that is, has a uniform spectral power density, then a reduction in bandwidth from B_W to B_N improves the S/N ratio by $20 \log(B_W/B_N)$ decibels.

2. Spectral Analysis

The role of the filter here is to permit observation of a narrow portion of a wideband spectrum. The selected band is specified by a center frequency f_0 and a bandwidth B, defined as the frequency interval about f_0 where the output/input ratio remains within 3 dB of that at the center. The bandwidth of the filter may be constant (i.e., independent of f_0) or a constant percentage of f_0.

The constant-bandwidth filter is advantageous in cases where the measured spectrum is rich in detail over a limited frequency range, for example, where a series of harmonics appears as the result of nonlinear distortion or where a number of sharp resonances are generated from a complex sound source. The constant-percentagebandwidth filter is more appropriate in cases where the measured spectrum encompasses a large number of decades, say two or more; where the source is unstable, constantly shifting its prominent frequencies; or where the power transmitted over a band of frequencies is of interest, as in noise control engineering.

Popular choices for the constant-percentage bandwidth are the octave (factor of 2), $\frac{1}{3}$, $\frac{1}{6}$, $\frac{1}{12}$, and $\frac{1}{24}$ octave. The bandwidth of a $\frac{1}{3}$-octave filter, for example, is $2^{1/6} f_0 - 2^{-1/6} f_0 = 0.231 f_0$. The $\frac{1}{3}$-octave filter, in fact, is the most widely used in acoustic spectral analysis. The reason is rooted in a property of human auditory response. Consider an experiment in which a human subject is exposed to a 60-dB narrow-band tone at 10 kHz. If the amplitude and center frequency of the tone remain fixed but the bandwidth increases, the subject will perceive no change in loudness until the bandwidth reaches 2.3 kHz, and then the loudness begins to increase. This is called the critical bandwidth and has a value of $\sim\frac{1}{3}$ octave. If the test is repeated at other, sufficiently high center frequencies, the resulting critical bandwidth remains at about $\frac{1}{3}$ octave. For sound measurements geared to human response, then, a narrower bandwidth does not influence loudness and a greater bandwidth yields a false measurement of loudness—hence, the choice of $\frac{1}{3}$-octave spectral resolution. A list of preferred $\frac{1}{3}$-octave center frequencies is given in Table IV. The audible spectrum, 20 Hz to 20 kHz, encompasses thirty-one $\frac{1}{3}$-octave bands.

TABLE IV Preferred $\frac{1}{3}$-Octave Center Frequencies[a]

16	20	25	31.5	40	50
63	80	100	125	160	200

[a] In hertz (also ×10 or ×100).

3. Spectral Shaping

The perceived loudness of a tone of constant amplitude is a strong function of frequency and amplitude. Many acoustic instruments feature not only a linear response, an objective measurement of sound pressure, but also a weighted response, which conforms to the frequency response of the human ear. The function of a weighting filter is to shape an acoustic spectrum to match the response of the ear. Three standard frequency response curves, called A, B, and C curves, conform to equal loudness curves at 40, 70, and 100 phons, respectively. A phon is a unit of loudness, usually specified in decibels; it is the same as the SPL at 1 kHz but differs at most other frequencies. The D weight has been proposed for applications involving aircraft noise measurement. The filter response curves for the A, B, C, and D weighting are shown in Fig. 17.

B. Spectrum Analyzers

A spectrum analyzer enables an observer to view the frequency spectrum of an acoustic time history on an output device such as a television monitor, chart recorder, or digital printer. A real-time analyzer produces a complete, continuously updated spectrum without interruption. The first real-time analyzers were analog in nature, based on either of two principles: (1) time compression, which used a frequency transformation to speed up processing time, or (2) a parallel bank of analog filters and detectors. The advent of VLSI (very large-scale integration) in the semiconductor industry made the all-digital, real-time analyzer a reality, offering competitive cost and enhanced stability, linearity, and flexibility.

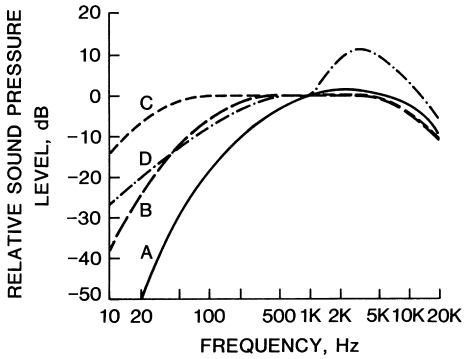


FIGURE 17 Response curves of A, B, C, and D weighting filters.

The spectrum analyzer performs the basic functions of preamplification, analog filtering, detection, analog-to-digital (A/D) conversion, logic control, computation, and output presentation. The frequency range usually covers the audio band but may exceed it at both ends. Digital real-time analyzers operate on either of two principles: the digital filter or the fast Fourier transform (FFT).

1. Digital Filter

The transfer function of a two-pole analog filter is written:

$$H(s) = \frac{(s + r_1)(s + r_2)}{(s + p_1)(s + p_2)} \tag{51}$$

where s is the Laplace operator, $r_{1.2}$ the zeros, and $p_{1.2}$ the poles. The filter characteristics—gain and cutoff frequencies—are fixed and can be changed only by changing the components making up the filter. The frequency response can be found by replacing s by $j\omega$.

The digital filter accepts samples $f(nT_s)$ of the time history from an A/D converter, where T_s is the sample interval, and yields an output in the form of a sequence of numbers. The transfer function is represented in the z domain:

$$H(z) = \frac{A_0 + A_1 z^{-1} + A_2 z^{-2}}{1 - B_1 z^{-1} - B_2 z^{-2}} \tag{52}$$

where $z^{-1} = \exp(-sT_s)$ is called the unit delay operator, since multiplication by z^{-1} is equivalent to delaying the sequence by one sample number. Synthesis of $H(z)$ requires a system that performs the basic operations of multiplying, summing, and delaying. Noteworthy is the fact that once the filter characteristics are set by choice of coefficients $A_0 \ldots B_2$, the frequency response parameters (center frequency f_0 and bandwidth B for a bandpass filter) are controlled by the sample rate $f_s = 1/T_s$. For example, doubling f_s doubles f_0 and B. Thus, the digital filter is a constant-percentage-bandwidth filter and is appropriate for those applications where such is required (Section II.A.2). Typically, the filters are six-pole Butterworth or Chebycheff filters of $\frac{1}{3}$-octave bandwidth. Several two-pole filters can be cascaded to produce filters of higher poles, or the data can be recirculated through the same filter several times.

2. Fast Fourier Transform

First consider the discrete Fourier transform (DFT), the digital version of Eq. (49),

$$F(k) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) \exp(-j2\pi kn/N) \tag{53}$$

where $f(n)$ is the value of the nth time sample, k the frequency component number, N the block size, or number of time samples. The time resolution depends on the time window $\Delta t = T/N$, and the frequency resolution depends on the sampling frequency $\Delta f = f_{\max}/N$. Obviously, the filter is a constant-bandwidth filter and again is suited to the appropriate applications (Section II.A.2). In contrast to the digital filtering technique, the data throughput is not continuous but is segmented into data blocks. Thus, for real-time analysis, the analyzer must be capable of processing one block of data while simultaneously acquiring a new block.

The FFT exploits the symmetry properties of the DFT to reduce the number of computations. The DFT requires N^2 multiplications to transform a data block of N samples from the time domain to the frequency domain: the FFT requires only $N \log_2 N$ multiplications. For a block size of $N = 1024$ samples, the reduction is over a factor of 100.

The DFT of Eq. (53) differs from the continuous Fourier transform in three ways, each presenting a data-processing problem that must be addressed by man/machine. First, the transformed function is a sampled time history. The sampling frequency must exceed twice the Nyquist frequency, as explained in Section II.A.1. In fact, it is beneficial to choose an even higher sampling frequency. For example, in a six-pole low-pass filter, the signal is down ~18 dB at $\frac{1}{2}$ octave past the cutoff frequency f_c. A strong component at this frequency will "fold over" as a component of frequency $\sqrt{2} f_c - f_c \approx 0.4 f_c$, attenuated only 18 dB, and may have a level comparable to the true signal. Increasing the sampling frequency to $f_s = 2.5 f_c$ will relieve the problem in this case.

Second, the filter time window yields the well-known $\sin x/x$ transform. In the frequency domain, the window spectrum is convolved with the signal spectrum and introduces ripples in the latter. The sidelobes of the $\sin x/x$ spectrum introduce leakage of power from a spectral component to its neighbors. A countermeasure to this effect is to use a Hanning window, a weighting time function that is maximum at the center of the window and zero at its edges. The Hanning window improves the sidelobe suppression at the expense of increased bandwidth. However, the Hanning window may not be needed if the signal is small at the edges of the window.

Finally, the digitally computed transform of the sampled time history must itself be presented as a sampled frequency spectrum. This fact is responsible for the so-called picket fence effect, whereby we do not observe the complete spectrum but only samples. Thus, we may miss a sharp peak and observe only the slopes. A Hanning window also helps to compensate for this effect.

Examples of acoustic signals necessitating analysis in real time are signals in the form of a sequence of transients,

as speech; aircraft flyover noise, as required by the Federal Aviation Administration (FAA); and measurements where the analyzer is an element in a control loop. For other types of signals, such as stationary or quasi-stationary signals, or transients shorter than the time window, the time history can be stored and analyzed at a later time.

3. Correlation

Many spectrum analyzers provide the capability of computing the autocorrelation and cross-correlation functions and their Fourier transforms, namely, the spectral and cross-spectral density functions. These operations are used to compare the data at one test station with that at another station. The cross-correlation of the time-varying functions $f_1(t)$ and $f_2(t)$ is expressed in terms of a time delay τ:

$$g_{12}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f_1(t) f_2(t+\tau)\, dt \qquad (54)$$

The Fourier transform of this function is the cross-power spectral density function:

$$G_{12}(f) = \int_{-\infty}^{\infty} g_{12}(\tau) \exp(-j2\pi f\tau)\, d\tau \qquad (55)$$

If $f_1(t)$ and $f_2(t)$ are the same signal, say at station 1, then Eqs. (54) and (55) yield $g_{11}(\tau)$ and $G_{11}(f)$, the autocorrelation function and the spectral power density function.

Two important acoustic applications of the cross-functions are transfer function determination and time delay estimation. Let us consider the transfer function. Suppose a noise or vibration source produces responses at two stations $f_1(t)$ and $f_2(t)$, having Fourier transforms $F_1(f)$ and $F_2(f)$. The transfer function $H_{12}(f) = F_2(f)/F_1(f)$ is related to the power spectra as follows:

$$H_{12}(f) = G_{12}(f)/G_{11}(f) \qquad (56)$$

Thus, Eq. (56) permits the determination of $H_{12}(f)$, while the source is operating in its natural condition.

Now consider time delay estimation. Suppose an acoustic signal propagates from station 1 to station 2 in time τ_0. Then $g_{12}(t)$ will show a peak at $\tau = \tau_0$, and $G_{12}(f)$ will have a phase angle $\phi_{12} = 2\pi f\tau_0$. If the signal is a pure tone, say a cosine wave, then $g_{12}(\tau)$ will also be a cosine wave of the same frequency but shifted by ϕ_{12}; that is, the maximum will be displaced by an angle ϕ_{12}. If the time delay τ_0 exceeds the period $1/f$ of the wave, then $g_{12}(\tau)$ will reveal two maxima and thus a twofold ambiguity in τ_0. Consequently, the maximum delay that can be uniquely determined is $\tau_{\max} < 1/f$. If the signal is a mixture of two tones of frequencies f_1 and f_2, then the maximum delay will be determined by the beat frequency, $\tau_{\max} < (f_2 - f_1)^{-1}$. Formal analysis leads to the criterion:

$$\tau_{\max} < 0.3/(f_2 - f_1) \qquad (57)$$

where $f_2 - f_1$ is the bandwidth of the signal. In these two cases, the cross-spectral density is strongly peaked at a few prominent frequencies. If, on the other hand, the time signal is strongly peaked as in the case of a narrow pulse, comprising a broad spectrum of frequencies, there is a criterion on minimum system bandwidth to measure a given delay similar to Eq. (57), with the inequality reversed.

The autocorrelation function reveals the presence of periodic signals in the presence of noise.

An important function in acoustic signal processing is the coherence function,

$$C_{12}(f) = \frac{|G_{12}(f)|^2}{G_{11}(f)G_{22}(f)} \qquad (58)$$

which has a value between 0 and 1. This function serves as a criterion as to whether the signals received at stations 1 and 2 have the same cause. It should have a reasonably high value even in measurement systems subject to noise and random events.

C. Sound Level Meters

A sound level meter is a compact portable instrument, usually battery-operated, for measuring SPL at a selected location. The microphone signal is preamplified (attenuated), weighted, again amplified (attenuated), detected, and displayed on an analog meter. The detector is a square-law detector followed by an averaging (mean or rms) network. There are a variety of additional features such as calibration, overload indication, and external connectors for filters and output signal.

The directional response of the microphone affects the accuracy of the measurement. In a free field, corrections are based on curves such as those in Fig. 7 if the angle of incidence is known. In a diffuse field, the random response curve must be relied on: The smaller the microphone, the more accurate are the results.

Two switch selections available to the user are weighting and time constant. The weighting networks are linear (unweighted), A, B, C, and sometimes D (Section II.A.3). For stationary or quasi-stationary signals, a "fast" or "slow" time constant, based on the response to a 200- or 500-msec signal, respectively, is used. The fast response follows time-varying sound pressures more closely at the expense of accuracy; the slow response offers a higher confidence level for the rms sound pressure measurement. Impulsive signals present something of a problem. Current standards specify a time constant of 35 msec, in an attempt to simulate the response of the human ear, plus the capability of storing the peak or rms value of the applied signal. To prevent saturation resulting from high peak amplitudes, the detector circuit must be capable of sustaining a crest factor, the ratio of peak to rms signal, of at least 5.

D. Storage of Acoustical Data

Up to the mid-1970s the workhorse of acoustical data storage was the magnetic tape recorder in both am and fm versions. The major limitation was the limited dynamic range, amounting to less than 40 dB for am tape and 50–55 dB for fm tape. This was followed by 7- to 9-track digital tape, which improved the dynamic range but in the 1980s yielded to VHS (video home systems) cassettes having greater storage density. Typical specifications for VHS cassette recorders, which are still on the market today, are 70 dB dynamic range, dc to 80-kHz frequency response, and recording time ranging from 50 min to 426.7 hr at sample rates of 1280 and 2.5 thousands of samples per second, respectively. With the explosive development of personal computers, the development of digital storage systems has proceeded at a comparable pace. These are classified as either *random access* or *sequential access* devices.

Random-access devices include hard drives, CD (compact disc) writers, and DVD (digital versatile disc) RAM (random-access memory) devices. The hard drive typically has a storage capacity of 20 gigabytes (GB) and a data transfer rate of over 10 megabytes (MB) per second. Traditional hard drives are not meant for archiving data nor for removal from one system to another. Now more options with removable hard-disk systems are available, such as the "Jaz" and "Orb," which have the disk in a removable cartridge. These cartridge-based hard drives have capacities of up to 2 GB and sustained data rates of over 8 MB/sec. The removable DVD-RAM has shown capacities of 5.2 GB and transfer rates of up to 1 MB per second.

While sequential access times are significantly greater than random-access devices, sequential access provides the highest storage capacities (up to 50 GB per tape) and very high sustained data transfer rates of over 6 MB/sec. The advent of advanced intelligent tape has an electronic memory device on each tape that speeds up the search process. In addition, 8-mm, digital linear tape, and 4-mm tapes are among forms of storage that allow up to 20 terabytes of information to be stored and accessed in a cost-effective manner.

High-quality digital storage devices conform to the Small Computer Systems Interface (SCSI) standard. The advantages are far-reaching. The conforming devices (including those mentioned above) are easily upgraded, mutually compatible, and interchangeable from one system to another. A single SCSI controller can control up to 15 independent SCSI devices.

An option available to users of digital storage devices is data compression, whereby data density is compressed by a two-to-one ratio. Most compression schemes are very robust and, combined with error detection and correction, produce error rates on the order of 10^{-15}.

E. The Computer as an Instrument in Acoustical Measurements

The integration of a digital computer into an acoustic measurement system offers many practical advantages in addition to improved specifications regarding dynamic range, data storage density, flexibility, and cost effectiveness. Many acoustic measurements require inordinately complex evaluation procedures. The capability of performing an on-line evaluation during a test provides the user with an immediate readout of the evaluated data: this may aid in the making of decisions regarding further data acquisition and the choice of test parameters. The decisionmaking procedure can even be automated. The digital data can readily be telecommunicated over ordinary telephone lines. Most digital systems accommodate a great variety of peripheral equipment.

Figure 18 shows examples of a computer integrated into an acoustical measurement system:

1. Active noise cancellation (Fig. 18a). The computer implements real-time digital filters 1 and 2, which serve as adaptive controllers to produce the required responses of noise-cancelling speakers 1 and 2.
2. Spatial transformation of sound fields (Fig. 18b). A cross spectrum analyzer yields a cross spectral representation of a sound field, based on acoustical measurements over a selected scan plane; then, a computer predicts the near field from the scan data using near field acoustic holography and the far field from the Helmholtz integral equation.
3. Computer-steered microphone arrays (Fig. 18c). In a large room, such as an auditorium or conference hall, the computer introduces a preprogrammed time delay in each microphone of a rectangular array, thus steering the array to the direction of high selectivity; coordinating more than one array, it controls the location from which the received sound is especially sensitive.

III. EXAMPLES OF ACOUSTICAL MEASUREMENTS

A. Measurement of Reverberation Time

Reverberation time (RT) is the time required for the sound in a room to decay over a specific dynamic range, usually taken to be 60 dB, when a source is suddenly interrupted. The Sabine formula relates the RT to the properties of the room.

$$T = 0.161V/\alpha S \tag{59}$$

where V is the volume of the room, S the area of its surfaces, and α the absorption coefficient due to losses

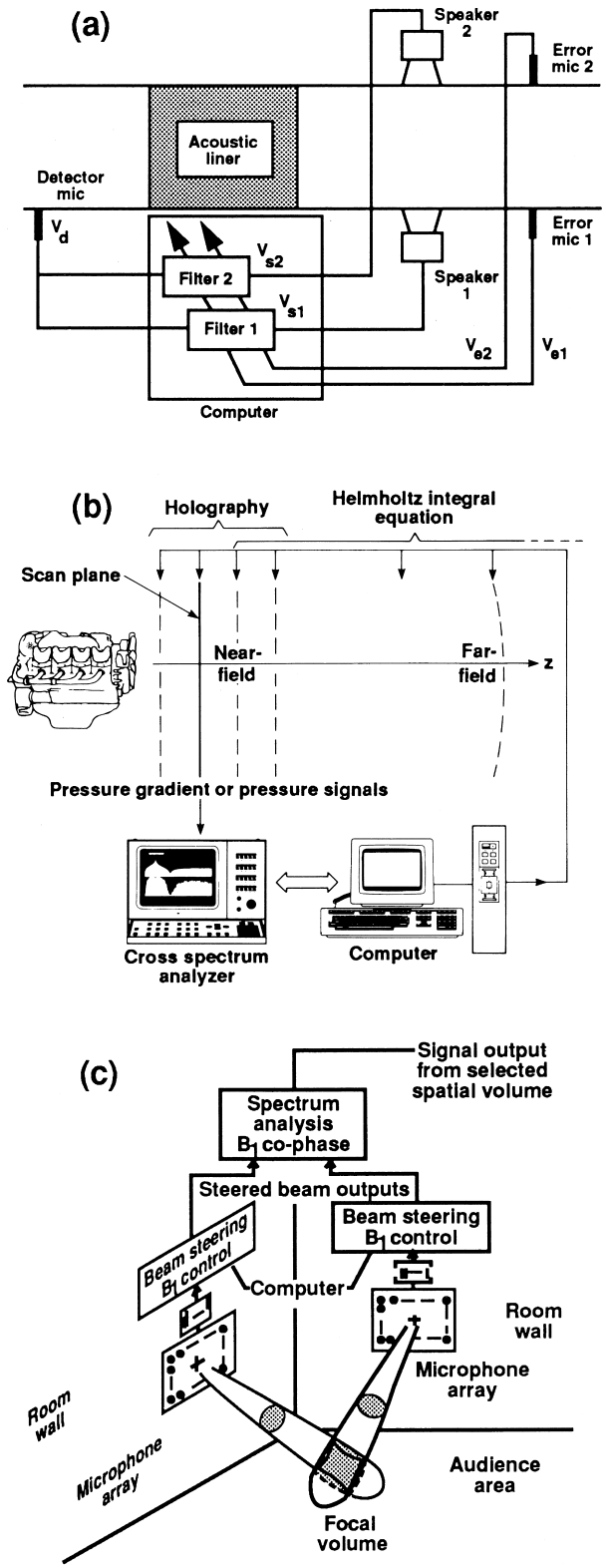


FIGURE 18 Measurement systems using computers. (a) Active noise cancellation. (b) Spatial transformation of sound fields. (c) Computer-steered microphone arrays. (Courtesy of NASA, B&K Instruments, and *J. Acoust. Soc. Am.*)

in the air and at the surfaces. Recommended values for a 500-Hz tone in a 1000-m^3 room are about 1.6 sec for a church, 1.2 sec for a concert hall, 1.0 sec for a broadcasting studio, and 0.8 sec for a motion picture theater, the values increasing slightly with room size. The room constant R, appearing in Eq. (32), is related to α through:

$$R = S\alpha/(1-\alpha) \tag{60}$$

A typical measuring arrangement is shown in Fig. 19a. A sound source is placed at a propitious location and the response is averaged over several microphone locations about the room.

If the source is excited into a pure tone, the measurement is beset with two basic difficulties. The act of switching generates additional tones, which establish beat frequencies and irregularities on the decay curves; furthermore, the excitation of room resonances can produce a break in the slope of the decay curve (Fig. 19b). The smoothness of the decay curve can be improved by widening the bandwidth of the source. Three types of excitation are used for this purpose: random noise, an impulse, or a warble tone, in which the center frequency is FM-modulated. The $\frac{1}{3}$-octave analyzer performs two function: It permits the frequency dependence of the RT to be determined, and it provides a logarithmic output to linearize the free decay curve. The output device can be a recorder (logarithmic if the analyzer provides a linear output) or a digital data acquisition system. The microphones can be multiplexed or measured individually. A typical decay curve obtained by this method is shown in Fig. 19c. Because many

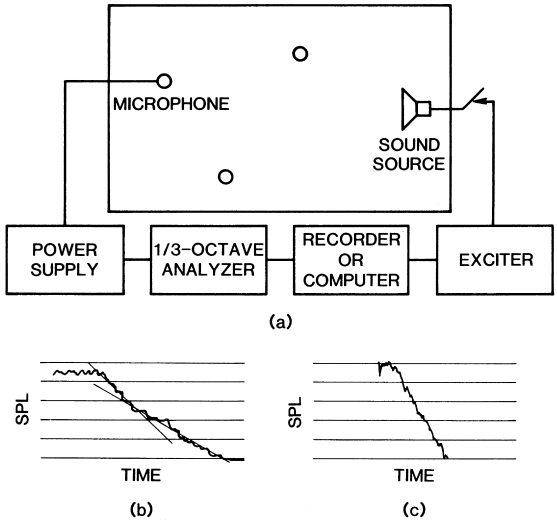


FIGURE 19 Measurement of reverberation time. (a) Experimental arrangement showing microphones (circles) positioned at suitable locations about the room. (b) Response curve showing a break in slope due to simultaneous room resonances. (c) Response curve showing unambiguous reverberation.

measurements are averaged to enhance confidence level, the method is time-consuming. If 20 averages are taken over each $\frac{1}{3}$-octave band from 125 Hz to 10 kHz, then 400 decay curves would have to be evaluated.

B. Measurement of Impulsive Noises

The measurement of noise from impulsive sources such as gunshots, explosives, punch presses, and impact hammers, as well as short transients in general, requires considerable care on the part of the observer. Sometimes these sources are under the observer's control, but on some occasions their occurrence is unpredictable, often affording but a single opportunity to make the measurement. By nature such sources are of large amplitude and short duration, requiring instruments capable of handling high crest factors and extended frequency content.

Measurement of the peak pressure does not give information on duration. Of greater interest is the measurement of rms pressure, from which loudness and energy content can be inferred. Such a measurement can be made with simple analog equipment, such as an impulse sound level meter. The pressure signal is squared and time-averaged, the square root extracted, and the result presented on a meter. The averaging time will affect the measurement. By convention, an averaging time constant of 35 msec is recommended in an effort to simulate the response of the human ear.

A description of an impulsive source in the frequency domain has several advantages. First, sources can be identified by their characteristic spectral signatures. Second, those components bearing a large amount of energy can be identified, as for noise control purposes. Finally, the response of an acoustic device to the signal is more readily analyzed in the frequency domain than in the time domain. Consider the Fourier spectrum of a pulse of constant amplitude A and duration T, shown in Fig. 20a:

$$F(f) = AT \sin(\pi f T)/\pi f T \tag{61}$$

Suppose the pulse is applied to an ideal, unity-gain filter of bandwidth B, center frequency f_0, and phase slope $t_L = d\phi/d\omega$. The spectrum of the pulse and transfer function of the filter are shown in Fig. 20b. The filter output will exhibit a characteristic ringing response; if $T \ll 1/B$, this can be approximated as:

$$\upsilon_0(t) \approx 2ABT \frac{\sin(\pi f_0 T)}{\pi f_0 T} \times \frac{\sin \pi B(t - t_L)}{\pi B(t - t_L)} \cos(2\pi f_0 t) \tag{62}$$

shown in Fig. 20c. The spectral component $F(f_0)$ is intimately related both to the peak response of the envelope, occurring at $t = t_L$, and to the integrated-squared response:

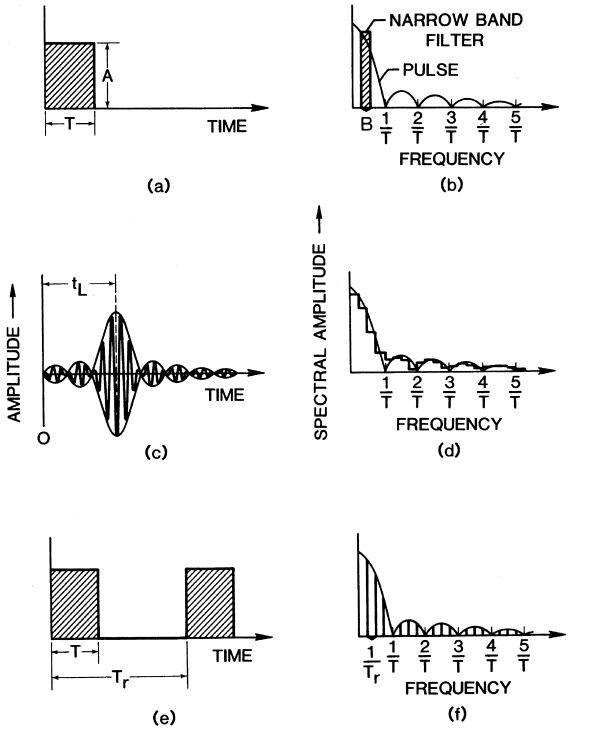


FIGURE 20 Measurement of impulsive noise. (a) Time history of a single pulse and (b) its amplitude–frequency spectrum together with that of an ideal narrow-band filter. (c) Time history of the filter response to the single pulse. (d) Reconstruction of the pulse spectrum from the outputs of several adjacent narrowband filters. (e) Time history of a periodic sequence of pulses and (f) its Fourier series amplitude spectrum (with envelope).

$$\upsilon_{0\,\text{peak}} \approx 2ABT \frac{\sin(\pi f_0 T)}{\pi f_0 T} = 2BF(f_0) \tag{63}$$

$$E = \int_{-\infty}^{\infty} \upsilon_0^2(t)\,dt \approx 2A^2BT^2 \left[\frac{\sin(\pi f_0 T)}{\pi f_0 T}\right]^2 = 2BF^2(f_0) \tag{64}$$

Thus, the Fourier spectrum can be reconstructed from the measurements corresponding to Eq. (63) or (64), using narrow-band filters of different center frequencies (Fig. 20d). If the condition $T \ll 1/B$ is not fulfilled, the filter response shows two bursts, each similar to that of Fig. 20c and separated by the pulse duration T, and Eqs. (63) and (64) are no longer valid.

If the impulsive noise is repetitive or if a single pulse can be reproduced repetitively, the pulse sequence can be represented as a Fourier series. The Fourier coefficients F_n are given by Eq. (61) if f is replaced by n/T_r, where T_r is the pulse repetition interval. The time history and Fourier spectrum are shown in Figs. 20e,f. The number of components per spectral lobe depends on the ratio T/T_r. Too large a ratio will yield too few components for

accurate representation of the spectrum, but too small a ratio must be avoided due to crest factor limitations in the analyzing equipment. A reasonable compromise is a ratio between 0.2 and 0.5. The reconstruction of short-transient spectra is a fine application of constant-bandwidth filtering.

Some final notes are pertinent to the measurement of transients:

1. The principles discussed here for the constant-amplitude pulse apply to short pulses of other shapes. The features of the spectrum of Fig. 20b are generally retained. In the case of a tone burst, the main lobe is displaced from the origin.
2. If the occurrence of the transient event cannot be predicted, a digital event recorder will prove useful. A time history is continuously sampled and transferred to a buffer. The buffer content is transferred to a storage register only when the signal exceeds a threshold. In this manner the pre- and postevent background signals are included on both sides of the event.
3. For long transients, such as sonic booms, a $\frac{1}{3}$-octave real-time analyzer may prove advantageous because the condition $T \ll 1/B$ may be difficult to fulfill, and energy content may spread over a wide band of frequencies. The rms response,

$$v_{0\,\mathrm{rms}} = \sqrt{E}/\tau \tag{65}$$

depends on the averaging (or integrating) time τ. For a fixed value of τ there will be a low-frequency rolloff due to the long response times, t_L, of the filters, which causes part of the signal to be excluded from the averaging. At high frequencies some error will occur because of high crest factors.

C. Measurement of Aircraft Noise

Aircraft noise measurements can be organized into two broad categories: aircraft noise monitoring and aircraft flyover testing, the latter for both engineering applications and noise certification.

1. Aircraft Noise Monitoring

Aircraft noise is measured routinely at numerous airports around the world to evaluate noise exposure in adjacent communities and to compare noise sources. The instruments are basically weather-protected sound level meters, covering an SPL range from about 60 to 120 dB at frequencies up to 10 kHz, and are generally installed near the airport boundaries.

2. Aircraft Flyover Testing for Certification

Federal Aviation Regulations, "Part 36—Noise Standards: Aircraft Type and Airworthiness Certification," define instrumentation requirements and test procedures for aircraft noise certification. The instrumentation system consists of microphones and their mounting, recording and reproducing equipment, calibrators, analysis equipment, and attenuators.

For subsonic transports and turbojet-powered airplanes, microphones are located on the extended centerline of the runway, 6500 m from the start of takeoff or 2000 m from the threshold of approach, and on the sideline 450 m from the runway. The microphones are of the capacitive type, either pressure or free field, with a minimum frequency response from 44–11,200 Hz. If the wind exceeds 6 knots, a windscreen is used.

If the recording and reproducing instrument is a magnetic tape recorder, it has a minimum dynamic range of 45 dB (noise floor to 3% distortion level), with a standard reading level 10 dB below the maximum and a frequency response comparable to that of the microphone.

The analyzer is a $\frac{1}{3}$-octave, real-time analyzer, having 24 bands in the frequency interval from 50–10,000 Hz. It has a minimum crest factor of 3, a minimum dynamic range of 60 dB, and a specified response time and provides an rms output from each filter every 500 msec.

Field calibrations are performed immediately before and after each day's testing. The microphone–preamplifier system is calibrated with a normal incidence pressure calibrator, the electronic system with "pink" noise (constant power in each $\frac{1}{3}$-octave band), and the magnetic tape recorder with the aid of a pistonphone.

After the recorded data are corrected to reference atmospheric conditions and reference flight conditions, an effective perceived noise level—a measure of subjective response—is evaluated.

The noise spectrum from a noncertification flyover of a Boeing 747 aircraft is shown in Fig. 21. The aircraft had

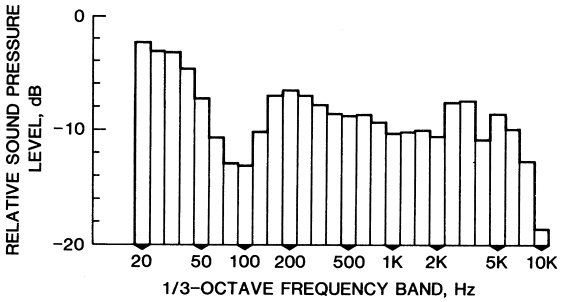


FIGURE 21 Noncertification flyover noise spectrum, in $\frac{1}{3}$ octaves, of a Boeing 747 aircraft. The reference level of 0 dB is arbitrary. (Courtesy of NASA.)

a speed of 130 m/sec, an altitude of 60 m, and a position directly over the microphone at the time the noise was recorded. The microphone was located 1.2 m above the ground, and the averaging time of the analyzer was 0.9 sec.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTIC CHAOS • ACOUSTICS, LINEAR • ACOUSTIC WAVE DEVICES • ANALOG SIGNAL ELECTRONIC CIRCUITS • SIGNAL PROCESSING, ACOUSTIC • ULTRASONICS AND ACOUSTICS • UNDERWATER ACOUSTICS

BIBLIOGRAPHY

Acoustical Society of America, Standards Secretariat, 120 Wall Street, 32nd Floor, New York, NY 10005-3993.

Crocker, M.J., ed.-in-chief (1997). "Encyclopedia of Acoustics," John Wiley & Sons, New York.

Hassall, J.R., and Zavari, K., (1979). "Acoustic Noise Measurement," 4th ed., Bruel & Kjaer Instruments, Marlborough, MA.

International Organization for Standardization (ISO), Case Postale 56, CH-1211, Geneve, Switzerland.

Kundert, W.R. (1978). *Sound and Vibration* **12,** 10–23.

Wong, G.S.K., and Embleton, T.F.W., eds. (1995). "AIP Handbook of Condenser Microphones," American Institute of Physics Press, New York.



Acoustics, Linear

Joshua E. Greenspon
J. G. Engineering Research Associates

I. Introduction
II. Physical Phenomena in Linear Acoustics
III. Basic Assumptions and Equations in Linear Acoustics
IV. Free Sound Propagation
V. Sound Propagation with Obstacles
VI. Free and Confined Waves
VII. Sound Radiation and Vibration
VIII. Coupling of Structure/Medium (Interactions)
IX. Random Linear Acoustics

GLOSSARY

Attenuation Reduction in amplitude of a wave as it travels.
Condensation Ratio of density change to static density.
Coupling Mutual interaction between two wave fields.
Diffraction Bending of waves around corners and over barriers.
Dispersion Dependence of velocity on frequency, manifested by distortion in the shape of a disturbance.
Elastic waves Traveling disturbances in solid materials.
Ergodic Statistical process in which each record is statistically equivalent to every other record. Ensemble averages over a large number of records at fixed times can be replaced by corresponding time averages on a single representative record.
Impedance Pressure per unit velocity.
Interaction Effect of two media on each other.
Medium Material through which a wave propagates.
Nondispersive medium Medium in which the velocity is independent of frequency and the shape of the disturbance remains undistorted.
Normal mode Shape function of a wave pattern in transmission.
Propagation Motion of a disturbance characteristic of radiaton or other phenomena governed by wave equations.
Ray Line drawn along the path that the sound travels, perpendicular to the wave front.
Reflection Process of a disturbance bouncing off an obstacle.
Refraction Change in propagation direction of a wave with change in medium density.
Reverberation Wave pattern set up in an enclosed space.
Scattering Property of waves in which a sound pattern is formed around an obstacle enveloped by an incoming wave.

Sommerfeld radiation condition Equation stating that waves must go out from their source towards infinity and not come in from infinity.
Standing waves Stationary wave pattern.
Wave guide Structure or channel along which the wave is confined.

ACOUSTICS is the science of sound—its generation, transmission, reception, and effects. Linear acoustics is the study of the physical phenomena of sound in which the ratio of density change to static density is small, typically much less than 0.1. A sound wave is a disturbance that produces vibrations of the medium in which it propagates.

I. INTRODUCTION

A unified treatment of the principles of linear acoustics must begin with the well-known phenomena of single-frequency acoustics. A second essential topic is random linear acoustics, a relatively new field, which is given a tutorial treatment in the final section of this article.

The objective is to present the elementary principles of linear acoustics and then to use straightforward mathematical development to describe some advanced concepts. Section II gives a physical description of phenomena in acoustics. Section III starts with the difference between linear and nonlinear acoustics and leads to the derivation of the basic wave equation of linear acoustics. Section IV discusses the fundamentals of normal-mode and ray acoustics, which is used extensively in studies of underwater sound propagation. In Section V, details are given on sound propagation as it is affected by barriers and obstacles. Sections VI–VIII deal with waves in confined spaces; sound radiation, with methods of solution to determine the sound radiated by structures; and the coupling of sound with its surroundings. Section IX discusses the fundamentals of radom systems as applied to structural acoustics.

II. PHYSICAL PHENOMENA IN LINEAR ACOUSTICS

A. Sound Propagation in Air, Water, and Solids

Many practical problems are associated with the propagation of sound waves in air or water. Sound does not propagate in free space but must have a dense medium to propagate. Thus, for example, when a sound wave is produced by a voice, the air particles in front of the mouth are vibrated, and this vibration, in turn, produces a disturbance in the adjacent air particles, and so on. [*See* ACOUSTICAL MEASUREMENT.]

If the wave travels in the same direction as the particles are being moved, it is called a longitudinal wave. This same phenomenon occurs whether the medium is air, water, or a solid. If the wave is moving perpendicular to the moving particles, it is called a transverse wave.

The rate at which a sound wave thins out, or attenuates, depends to a large extent on the medium through which it is propagating. For example, sound attenuates more rapidly in air than in water, which is the reason that sonar is used more extensively under water than in air. Conversely, radar (electromagnetic energy) attenuates much less in air than in water, so that it is more useful as a communication tool in air.

Sound waves travel in solid or fluid materials by elastic deformation of the material, which is called an elastic wave. In air (below a frequency of 20 kHz) and in water, a sound wave travels at constant speed without its shape being distorted. In solid material, the velocity of the wave changes, and the disturbance changes shape as it travels. This phenomenon in solids is called dispersion. Air and water are for the most part nondispersive media, whereas most solids are dispersive media.

B. Reflection, Refraction, Diffraction, Interference, and Scattering

Sound propagates undisturbed in a nondispersive medium until it reaches some obstacle. The obstacle, which can be a density change in the medium or a physical object, distorts the sound wave in various ways. (It is interesting to note that sound and light have many propagation characteristics in common: The phenomena of reflection, refraction, diffraction, interference, and scattering for sound are very similar to the phenomena for light.) [*See* WAVE PHENOMENA.]

1. Reflection

When sound impinges on a rigid or elastic obstacle, part of it bounces off the obstacle, a characteristic that is called reflection. The reflection of sound back toward its source is called an echo. Echoes are used in sonar to locate objects under water. Most people have experienced echoes in air by calling out in an empty hall and hearing their words repeated as the sound bounces off the walls.

2. Refraction and Transmission

Refraction is the change of direction of a wave when it travels from a medium in which it has one velocity to a medium in which it has a different velocity. Refraction of sound occurs in the ocean because the temperature or the water changes with depth, which causes the velocity of

sound also to change with depth. For simple ocean models, the layers of water at different temperatures act as though they are layers of different media. The following example explains refraction: Imagine a sound wave that is constant over a plane (i.e., a plane wave) in a given medium and a line drawn perpendicular to this plane (i.e., the normal to the plane) which indicates the travel direction of the wave. When the wave travels to a different medium, the normal bends, thus changing the direction of the sound wave. This normal line is called a *ray* and is discussed later with ray acoustics in Section IV.A.

When a sound wave impinges on a plate, part of the wave reflects and part goes through the plate. The part that goes through the plate is the transmitted wave. Reflection and transmission are related phenomena that are used extensively to describe the characteristics of sound baffles and absorbers.

3. Diffraction

Diffraction is associated with the bending of sound waves around or over barriers. A sound wave can often be heard on the other side of a barrier even if the listener cannot see the source of the sound. However, the barrier projects a shadow, called the shadow zone, within which the sound cannot be heard. This phenomenon is similar to that of a light that is blocked by a barrier.

4. Interference

Interference is the phenomenon that occurs when two sound waves converge. In linear acoustics the sound waves can be superimposed. When this occurs, the waves interfere with each other, and the resultant sound is the sum of the two waves, taking into consideration the magnitude and the phase of each wave.

5. Scattering

Sound scattering is related closely to reflection and transmission. It is the phenomenon that occurs when a sound wave envelops an obstacle and breaks up, producing a sound pattern around the obstacle. The sound travels off in all directions around the obstacle. The sound that travels back toward the source is called the backscattered sound, and the sound that travels away from the source is known as the forwardscattered field.

C. Standing Waves, Propagating Waves, and Reverberation

When a sound wave travels freely in a medium without obstacles, it continues to propagate unless it is attentuated by some characteristic of the medium, such as absorption. When sound waves propagate in an enclosed space, they reflect from the walls of the enclosure and travel in a different direction until they hit another wall. In a regular enclosure, such as a rectangular room, the waves reflect back and forth between the sound source and the wall, setting up a constant wave pattern that no longer shows the characteristics of a traveling wave. This wave pattern, called a standing wave, results from the superposition of two traveling waves propagating in opposite directions. The standing wave pattern exists as long as the source continues to emit sound waves. The continuous rebounding of the sound waves causes a reverberant field to be set up in the enclosure. If the walls of the enclosure are absorbent, the reverberant field is decreased. If the sound source stops emitting the waves, the reverberant standing wave field dies out because of the absorptive character of the walls. The time it takes for the reverberant field to decay is sometimes called the time constant of the room.

D. Sound Radiation

The interaction of a vibrating structure with a medium produces disturbances in the medium that propagate out from the structure. The sound field set up by these propagating disturbances is known as the sound radiation field. Whenever there is a disturbance in a sound medium, the waves propagate out from the disturbance, forming a radiation field.

E. Coupling and Interaction between Structures and the Surrounding Medium

A structure vibrating in air produces sound waves, which propagate out into the air. If the same vibrating structure is put into a vacuum, no sound is produced. However, whether the vibrating body is in a vacuum or air makes little difference in the vibration patterns, and the reaction of the structure to the medium is small. If the same vibrating body is put into water, the high density of water compared with air produces marked changes in the vibration and consequent radiation from the structure. The water, or any heavy liquid, produces two main effects on the structure. The first is an added mass effect, and the second is a damping effect known as radiation damping. The same type of phenomenon also occurs in air, but to a much smaller degree unless the body is traveling at high speed. The coupling phenomenon in air at these speeds is associated with flutter.

F. Deterministic (Single-Frequency) Versus Random Linear Acoustics

When the vibrations are not single frequency but are random, new concepts must be introduced. Instead of dealing

with ordinary parameters such as pressure and velocity, it is necessary to use statistical concepts such as auto- and cross-correlation of pressure in the time domain and auto- and cross-spectrum of pressure in the frequency domain. Frequency is a continuous variable in random systems, as opposed to a discrete variable in single-frequency systems. In some acoustic problems there is randomness in both space and time. Thus, statistical concepts have to be applied to both time and spatial variables.

III. BASIC ASSUMPTIONS AND EQUATIONS IN LINEAR ACOUSTICS

A. Linear Versus Nonlinear Acoustics

The basic difference between linear and nonlinear acoustics is determined by the amplitude of the sound. The amplitude is dependent on a parameter, called the condensation, that describes how much the medium is compressed as the sound wave moves. When the condensation reaches certain levels, the sound becomes nonlinear. The major difference between linear and nonlinear acoustics can best be understood by deriving the one-dimensional wave equation for sound waves and studying the parameters involved in the derivation. Consider a plane sound wave traveling down a tube, as shown in Fig. 1.

Let the cross-sectional area of the tube be A and let ξ be the particle displacement along the x axis from the equilibrium position. Applying the principle of conservation of mass to the volume $A\,dx$ before and after it is displaced, the following equation is obtained:

$$\rho A\,dx(1+\partial\xi/\partial x)=\rho_o A\,dx \tag{1}$$

The mass of the element before the disturbance arrives is $\rho_o A\,dx$ where ρ_o is the original density of the medium. The mass of this element as the disturbance passes is changed to:

$$\rho A\,dx(1+\partial\xi/\partial x)$$

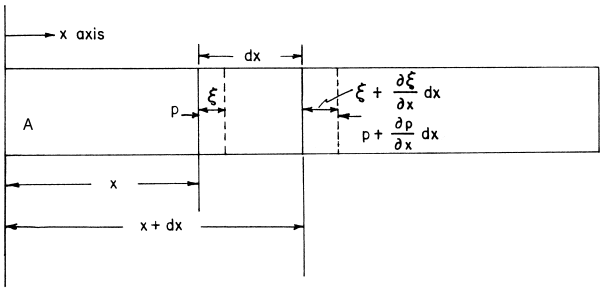


FIGURE 1 Propagation of a plane one-dimensional sound wave. A = cross sectional area of tube; ξ = particle displacement along the x axis; p = acoustic pressure.

where ρ is the new density of the disturbed medium. This disturbed density ρ can be defined in terms of the original density ρ_o by the following relation:

$$\rho=\rho_o(1+S) \tag{2}$$

where S is called the condensation. By substituting Eq. (2) into (1) we obtain:

$$(1+S)(1+\partial\xi/\partial x)=1 \tag{3}$$

If p is the sound pressure at x, then $p+\partial p/\partial x\,dx$ is the sound pressure at $x+dx$ (by expanding p into a Taylor series in x and neglecting higher-order terms in dx). Applying Newton's law to the differential element, we find that:

$$-\frac{\partial p}{\partial x}=\rho_o\frac{\partial^2\xi}{\partial t^2} \tag{4}$$

If it is assumed that the process of sound propagation is adiabatic (i.e., there is no change of heat during the process), then the pressure and density are related by the following equation:

$$\frac{P}{p_o}=\left(\frac{\rho}{\rho_o}\right)^{\gamma} \tag{5}$$

where $P=$ total pressure $=p+p_o$, p is the disturbance sound pressure, and γ is the adiabatic constant, which has a value of about 1.4 for air. Using Eqs. (2) and (3) gives:

$$\rho=\frac{\rho_o}{1+\partial\xi/\partial x}$$

Thus,

$$\frac{\partial p}{\partial x}=\frac{\partial P}{\partial x}=-\gamma p_o\left(1+\frac{\partial\xi}{\partial x}\right)^{-1-\gamma}\frac{\partial^2\xi}{\partial x^2} \tag{6}$$

Substituting into Eq. (4) gives:

$$\gamma p_o\frac{\partial^2\xi/\partial x^2}{(1+\partial\xi/\partial x)^{1+\gamma}}=\rho_o\frac{\partial^2\xi}{\partial t^2} \tag{7}$$

or finally,

$$c^2\frac{\partial^2\xi/\partial x^2}{(1+\partial\xi/\partial x)^{1+\gamma}}=\frac{\partial^2\xi}{\partial t^2} \tag{8}$$

where $c^2=p_o\gamma/\rho_o$ (c is the sound speed in the medium). If $\partial\xi/\partial x$ is small compared with 1, then Eq. (3) gives:

$$S=-\frac{\partial\xi}{\partial x} \tag{9}$$

and (8) gives:

$$c^2\frac{\partial^2\xi}{\partial x^2}=\frac{\partial^2\xi}{\partial t^2} \tag{10}$$

Thus,

$$\xi=f_1(x-ct)+f_2(x-ct) \tag{10a}$$

Equations (9) and (10) are the linear acoustic approximations. The first term in Eq. (10a) is an undistorted traveling wave that is moving at speed c in the $+x$ direction, and the second term is an undistorted traveling wave moving with speed c in the $-x$ direction.

Condensation values S of the order of 1 are characteristic of sound waves with amplitudes approaching 200 db rel. 0.0002 dyne/cm^2. The threshold of pain is about 130 db rel. 0.0002 dyne/cm^2. This is the sound pressure level that results in a feeling of pain in the ear. The condensation value for this pressure is about $S = 0.0001$. For a condensation value $S = 0.1$, we are in the nonlinear region. This condensation value corresponds to a sound pressure level of about 177 db rel. 0.0002 dyne/cm^2. All the ordinary sounds that we hear such as speech and music (even very loud music) are usually well below 120 db rel. 0.0002 dyne/cm^2. A person who is very close to an explosion or is exposed to sonar transducer sounds underwater would suffer permanent damage to his hearing because the sounds are usually well above 130 db rel. 0.0002 dyne/cm^2.

B. Derivation of Basic Equations

It is now necessary to derive the general three-dimensional acoustic wave equations. In Section III.A, the one-dimensional wave equation was derived for both the linear and nonlinear cases. If there is a fluid particle in the medium with coordinates x, y, z, the fluid particle can move in three dimensions. Let the displacement vector of the particle be **b** having components ξ, η, ζ, as shown in Fig. 2.

The velocity vector **q** is

$$\mathbf{q} = \partial \mathbf{b}/\partial t \tag{11}$$

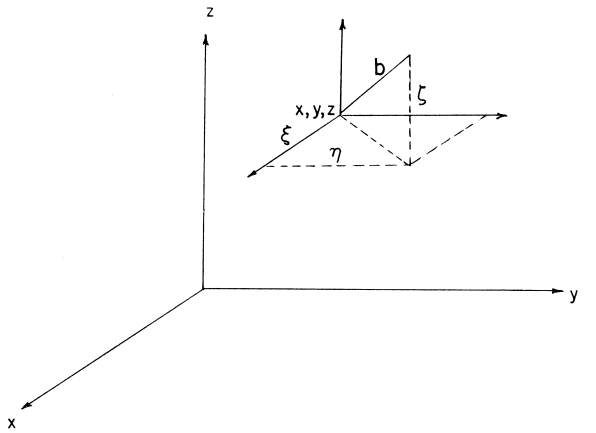


FIGURE 2 The fluid particle. x, y, z = rectangular coordinates of fluid particle; **b** = displacement vector of the particle (components of **b** are ξ, η, ζ).

Let this velocity vector have components u, v, w where

$$u = \partial\xi/\partial t \qquad v = \partial\eta/\partial t \qquad w = \partial\zeta/\partial t \tag{12}$$

As the sound wave passes an element of volume, $V = dx\,dy\,dz$, the element changes volume because of the displacement ξ, η, ζ. The change in length of the element in the x, y, z directions, respectively, is $(\partial\xi/\partial x)\,dx$, $(\partial\eta/\partial y)\,dy$, $(\partial\zeta/\partial z)\,dz$; so the new volume is $V + \Delta V$ where:

$$V + \Delta V = dx\left(1 + \frac{\partial\xi}{\partial x}\right) dy\left(1 + \frac{\partial\eta}{\partial y}\right) dz\left(1 + \frac{\partial\rho}{\partial z}\right) \tag{13}$$

The density of the medium before displacement is ρ_o and the density during displacement is $\rho_o(1+S)$, as in the one-dimensional case developed in the last section. Applying the principle of conservation of mass to the element before and after displacement, we find that:

$$(1 + S)(1 + \partial\xi/\partial x)(1 + \partial\eta/\partial y)(1 + \partial\zeta/\partial z) = 1 \tag{14}$$

Now we make the linear acoustic approximation that $\partial\xi/\partial x$, $\partial\eta/\partial y$, and $\partial\zeta/\partial z$ are small compared with 1. So Eq. (14) becomes the counterpart of Eq. (9) in one dimension:

$$S = -(\partial\xi/\partial x + \partial\eta/\partial y + \partial\zeta/\partial z) \tag{15}$$

This equation is called the equation of continuity for linear acoustics.

The equations of motion for the element $dx\,dy\,dz$ are merely three equations in the three coordinate directions that parallel the one-dimensional Eq. (4); thus, the three equations of motion are

$$-\frac{\partial p}{\partial x} = \rho_o\frac{\partial^2\xi}{\partial t^2} \qquad -\frac{\partial p}{\partial y} = \rho_o\frac{\partial^2 y}{\partial t^2} \qquad -\frac{\partial p}{\partial z} = \rho_o\frac{\partial^2\rho}{\partial t^2} \tag{16}$$

If one differentiates the first of these equations with respect to x, the second with respect to y, and the third with respect to z, and adds them, then letting

$$\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$$

one obtains:

$$\nabla^2 p = \rho_o\frac{\partial^2 S}{\partial t^2} \tag{17}$$

Now we introduce the adiabatic assumption in Eq. (17); that is,

$$\frac{P}{p_o} = \left(\frac{\rho}{\rho_o}\right)^\gamma \tag{18}$$

where $P =$ total pressure $= p + p_o$ and p is the sound pressure due to the disturbance. Since

$$\begin{aligned} \rho &= \rho_o(1 + S), \\ P/p_o &= (1 + S)^\gamma \end{aligned} \tag{19}$$

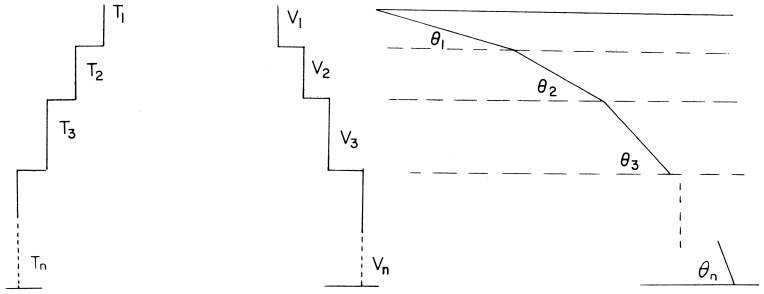


FIGURE 3 Temperatures, velocities, and refraction angles of the sound. $T_1, T_2, \ldots, T_n$ = temperatures of the n layers of the model medium; $V_1, V_2, \ldots, V_n$ = sound velocities in the n layers of the model medium.

For small S, the binomial theorem applied to Eq. (19) gives:

$$\frac{P - p_o}{\rho_o} = Sc^2 \tag{20}$$

(c being the adiabatic sound velocity, as discussed for the one-dimensional case). Thus,

$$p = \rho_o S c^2$$

Substituting into Eq. (17) we obtain:

$$c^2 \nabla^2 p = \partial^2 p / \partial t^2 \tag{21}$$

C. Intensity and Energy

The one-dimensional equation for place waves is given by Eq. (10). The displacement for a harmonic wave can be written:

$$\xi = Ae^{i(\omega t + kx)}$$

The pressure is given by Eq. (20); that is, $p = \rho_o C_o^2 S$, where $S = \partial \xi / \partial x$ for the one-dimensional wave. Then,

$$p = \rho_o c_o^2 \frac{\partial \xi}{\partial x}$$

The velocity is given by $u = \partial \xi / \partial t$, so, for one-dimensional harmonic waves, $p = \rho_o c_o^2 (ik)$ and $u = i\omega\xi$, but $k = \omega / c_o$. Thus, $p = \rho_o c_o u$. The intensity is defined as the power flow per unit area (or the rate at which energy is transmitted per unit area). Thus, $I = p\xi$. The energy per unit area is the work done on the medium by the pressure in going through displacement ξ, that is, $E_f = p\dot{\xi}$. And by the above,

$$I = p^2 / \rho_o c_o$$

IV. FREE SOUND PROPAGATION

A. Ray Acoustics

Characteristics of sound waves can be studied by the same theory regardless of whether the sound is propagating in air or water. The simplest of the sound-propagation theories is known as ray acoustics. A sound ray is a line drawn normal to the wave front of the sound wave as the sound travels. In Section II. B. 2, refraction was described as the bending of sound waves when going from one medium to another. When a medium such as air or water has a temperature gradient, then it can be thought of as having layers that act as different media. The objective of this theory, or any other transmission theory, is to describe the sound field at a distance from the source. There are two main equations of ray theory. The first is Snell's law of refraction, which states that

$$\frac{V_1}{\cos\theta_1} = \frac{V_2}{\cos\theta_2} = \frac{V_3}{\cos\theta_3} = \cdots = \frac{V_n}{\cos\theta_n} \tag{22}$$

where $V_1, V_2, \ldots V_n$ are the velocities of sound through the various layers of the medium, which are at different temperatures as shown in Fig. 3.

The second relation is that the power flow remains constant along a ray tube (i.e., there is conservation of energy along a ray tube). A ray tube is a closed surface formed by adjacent rays, as shown in Fig. 4. If the power flow remains constant along a ray tube, then,

$$\frac{p_1^2 A_1}{\rho_o c_1} = \frac{p_2^2 A_2}{\rho_o c_2} = \cdots \frac{p_n^2 A_n}{\rho_o c_n} \tag{23}$$

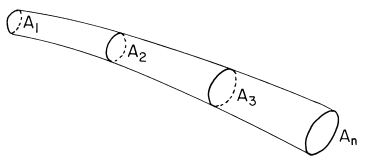


FIGURE 4 Ray tube, $A_1, A_2, \ldots, A_n$ = cross section area of the ray tube at the n stations along the tube.

where p refers to the sound pressure, A is the cross-sectional area of the ray tube, ρ_o the mass density of the medium, and C the sound velocity in the medium. But, $p^2/\rho_o c = I$, the sound intensity. Thus,

$$I_1 A_1 = I_2 A_2 = \cdots I_n A_n \tag{24}$$

The intensity can therefore be found at any point if the initial intensity I_1 and the areas of the ray tube $A_1, \ldots, A_n$ are known. The ray tube and the consequent areas can be determined by tracing the rays. The tracing is done by using Snell's law (Eq. 22). The velocities of propagation $V_1, V_2, \ldots, V_n$ are determined from the temperature and salinity of the layer. One such equation for sound velocity is

$$\begin{aligned} V = {} & 1449 + 4.6T - 0.055T^2 + 0.0003T^3 \\ & + (1.39 - 0.012T)(s - 35) + .017d \end{aligned} \tag{25}$$

where V is the velocity of sound in meters per second, T the temperature in degrees centigrade, s the salinity in parts per thousand, and d the depth in meters. The smaller the ray tubes, that is, the closer together the rays, the more accurate are the results.

Simple ray-acoustics theory is good only at high frequencies (usually in the kilohertz region). For low frequencies (e.g., less than 100 Hz), another theory, the normal mode theory, has to be used to compute transmission characteristics.

B. Normal Mode Theory

The normal mode theory consists of forming a solution of the acoustic wave equation that satisfies specific boundary conditions. Consider the sound velocity $C(z)$ as a function of the depth z, and let h be the depth of the medium. The medium is bounded by a free surface at $z = 0$ and a rigid bottom at $z = h$. Let a point source be located at $z = z_1$, $r = 0$ (in cylindrical coordinates) as shown in Fig. 5.

The pressure p is given by the Helmholtz equation:

$$\begin{aligned} \frac{\partial^2 p}{\partial r^2} + \frac{1}{r}\frac{\partial p}{\partial r} + \frac{\partial^2 p}{\partial z^2} + k^2(z)p &= -\frac{2}{r}\delta(z - z_1)\delta(r) \\ k^2(z) &= \frac{\omega^2}{c(z)^2} \end{aligned} \tag{26}$$

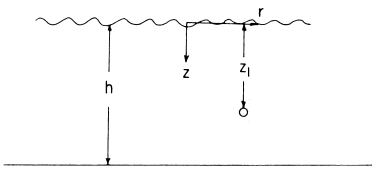


FIGURE 5 Geometry for normal mode propagation. r, z = cylindrical coordinates of a point in the medium; h = depth of medium; $z_1 = z$ coordinate of the source.

The δ functions describe the point source. The boundary conditions are

$$\begin{aligned} p(r, o) &= 0 \qquad \text{(free surface)} \\ \frac{\partial p}{\partial z}(r, h) &= 0 \qquad \text{(rigid bottom)} \end{aligned} \tag{27}$$

Equations (26) and (27) essentially constitute all of the physics of the solution. The rest is mathematics. The solution of the homogeneous form of Eq. (26) is first found by separation of variables. Since the wave has to be an outgoing wave, this solution is

$$p(r, z) = H_o^{(1)}(\xi r)\psi(z, \xi) \tag{28}$$

where $H_o^{(1)}$ is the Hankel function of the first kind of order zero. The function $\psi(z, \xi)$ satisfies the equation:

$$\frac{d^2\psi}{dz^2} + [k^2(z) - \xi^2]\psi = 0 \tag{29}$$

with boundary conditions:

$$\psi(o, \xi) = 0 \qquad \frac{d\psi(h, \xi)}{dz} = 0 \tag{30}$$

Since Eq. (30) is a second-order linear differential equation, let the two linearly independent solutions be $\psi_1(z, \xi)$ and $\psi_2(z, \xi)$. Thus, the complete solution is

$$\psi(z, \xi) = B_1\psi_1(z, \xi) + B_2\psi_2(z, \xi) \tag{31}$$

where B_1 and B_2 are constants.

Substitution of Eq. (31) into (30) leads to an equation from which the allowable values of ξ (the eigenvalues) can be obtained, that is,

$$\psi_1(o, \xi)\frac{d\psi_2(h, \xi)}{dz} - \psi_2(o, \xi)\frac{d\psi_1(h, \xi)}{dz} = 0 \tag{32}$$

The nth root of this equation is called ξ_n. The ratio of the constants B_1 and B_2 is

$$\frac{B_1}{B_2} = -\frac{\psi_2(o, \xi_n)}{\psi_1(o, \xi_n)} \tag{33}$$

The $H_o^{(1)}(\xi_n r)\psi(z, \xi_n)$ are known as the normal mode functions, and the solution of the original inhomogeneous equation can be expanded in terms of these normal mode functions as follows:

$$p(r, z) = \sum_n A_n H_o^{(1)}(\xi_n r)\psi(z, \xi_n) \tag{34}$$

with unknown coefficients A_n, which will be determined next. Substituting Eq. (34) into (26) and employing the relation for the Hankel function,

$$\left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr} + \xi_n^2\right)H_o^{(1)}(\xi_n r) = \frac{2i}{\pi r}\delta(r) \tag{35}$$

leads to:

$$\sum_n A_n \psi_n(z) = i\pi\delta(z - z_1) \tag{36}$$

Next one must multiply Eq. (36) by $\psi_m(z)$ and integrate over the depth 0 to h. Using the orthogonality of mode assumption, which states:

$$\int_0^h \psi_n(z)\psi_m(z)\,dz = 0 \qquad \text{if } m \neq n \tag{37}$$

we find that:

$$A_n = \frac{i\pi\psi_n(z_1)}{\int_0^h \psi_n^2(z)\,dz} \tag{38}$$

So,

$$p(r, z) = \pi i \sum_n \psi_n(z_1)\psi_n(z)\frac{H_o^{(1)}(\xi_n r)}{\int_0^h \psi_n^2(z)\,dz} \tag{39}$$

If the medium consists of a single layer with constant velocity, C_o, it is found that:

$$\begin{aligned} \psi_n(z) &= \cosh b_n z & \xi_n &= \sqrt{b_n^2 + k^2} \\ b_n &= i\left(n + \tfrac{1}{2}\right)\pi/h & k &= \omega/c_o \end{aligned} \tag{40}$$

C. Underwater Sound Propagation

Ray theory and normal mode theory are used extensively in studying the transmission of sound in the ocean. At frequencies below 100 Hz, the normal mode theory is necessary.

Two types of sonar are used underwater: active and passive. Active sonar produces sound waves that are sent out to locate objects by receiving echos. Passive sonar listens for sounds. Since the sound rays are bent by refraction (as discussed in Section IV.B), there are shadow zones in which the sound does not travel. Thus a submarine located in a shadow zone has very little chance of being detected by sonar.

Since sound is the principal means of detection under water, there has been extensive research in various aspects of this field. The research has been divided essentially into three areas: generation, propagation, and signal processing. Generation deals with the mechanisms of producing the sound, propagation deals with the transmission from the source to the receiver, and signal processing deals with analyzing the signal to extract information.

D. Atmospheric Sound Propagation

It has been shown that large amplitude sounds such as sonic booms from supersonic aircraft can be detected at very low frequencies (called infrasonic frequencies) at distances above 100 km from the source. In particular, the Concorde sonic boom has been studied at about 300 km, and signals of about 0.6 N/m^2 were received at frequencies of the order of 0.4 Hz. The same phenomenon occurs for thunder and explosions on the ground.

The same principles hold in the atmosphere as in water for the bending of rays in areas of changing temperature. Because of the large attenuation of higher frequency sound in air as opposed to water, sound energy is not used for communication in air. For example, considering the various mechanisms of absorption in the atmosphere, the total attenuation is about 24 db per kiloyard at 100 Hz, whereas, for underwater, the sound attenuates at 100 Hz at about 0.001 db per kiloyard.

V. SOUND PROPAGATION WITH OBSTACLES

A. Refraction

Refraction is the bending of sound waves. (Section II.B.2.) The transmission of sound through the water with various temperature layers and the application of Snell's law have already been treated in this article. Transmission of sound through water is probably the most extensive practical application of refraction in acoustics.

B. Reflection and Transmission

There are many practical problems related to reflection and transmission of sound waves. One example is used here to acquaint the reader with the concepts involved in the reflection and transmission of sound.

Consider a sound wave coming from one medium and hitting another, as shown in Fig. 6. What happens in layered media, such as the temperature layers described in connection with ray acoustics and underwater sound, can now be noted. When ray acoustics and underwater sound were discussed, only refraction and transmission were described. The entire process for one transition layer can now be explained.

The mass density and sound velocity in the upper medium is ρ, c and in the lower medium is ρ_1, c_1. The pressure in the incident wave, p_{inc}, can be written:

$$p_{inc} = p_o e^{ik(x\sin\theta - z\cos\theta)} \qquad k = \omega/c \tag{41}$$

(i.e., assuming a plane wave front).

From Snell's law, it is found that the refracted wave angle θ_1 is determined by the relation:

$$\frac{\sin\theta}{\sin\theta_1} = \frac{c}{c_1} \tag{42}$$

There is also a reflected wave that goes off at angle θ, as shown in the figure. The magnitude of the reflected wave

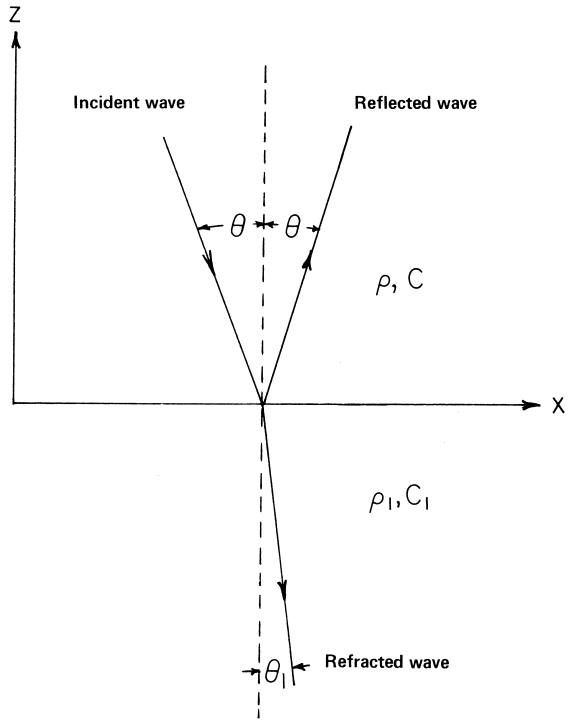


FIGURE 6 Reflection and transmission from an interface. ρ, c = mass density and sound velocity in upper medium; ρ_1, c_1 = mass density and sound velocity in lower medium; θ = angle of incidence and reflection; θ_1 = angle of refraction.

is not known, but since its direction θ is known, it can be written in the form:

$$p_{\text{refl}} = Ve^{ik(x\sin\theta + z\cos\theta)} \tag{43}$$

where V is the reflection coefficient. Similarly, the refracted wave can be written in the form

$$p_{\text{refrac}} = Wp_o e^{ik_1(x\sin\theta_1 - z\cos\theta_1)} \tag{44}$$

where $k_1 = \omega/c_1$ and W is the transmission coefficient. The boundary conditions at $z = 0$ are

$$\begin{aligned} p_{\text{upper}} &= p_{\text{lower}} && \text{(acoustic pressure is continuous across the boundary)} \\ (v_z)_{\text{upper}} &= (v_z)_{\text{lower}} && \text{(particle velocity normal to boundary is continuous across the boundary)} \end{aligned} \tag{45}$$

The velocity is related to the pressure by the expression:

$$\frac{\partial p}{\partial z} = \rho\frac{\partial v_z}{\partial t} \tag{46}$$

For harmonic motion $v \sim e^{i\omega t}$, so:

$$\partial p/\partial z = i\omega\rho v_z \tag{47}$$

The second boundary condition at $z = 0$ is, therefore,

$$\frac{1}{\rho}\frac{\partial p_{\text{upper}}}{\partial z} = \frac{1}{\rho_1}\frac{\partial p_{\text{lower}}}{\partial z} \tag{48}$$

The total field in the upper medium consists of the reflected and incident waves combined, so:

$$\begin{aligned} p_{\text{upper}} &= p_{\text{inc}} + p_{\text{refl}} \\ &= p_o e^{ikx\sin\theta}(e^{-ikz\cos\theta} + Ve^{ikz\cos\theta}) \end{aligned} \tag{49}$$

Substituting into the boundary conditions, we find that:

$$p_o e^{ikx\sin\theta}(1 + V) = Wp_o e^{ik_1 x\sin\theta_1}$$

so

$$1 + V = We^{ik_1 x\sin\theta_1 - ikx\sin\theta} \tag{50}$$

Since $1 + V$ is independent of x, then $e^{ik_1 x\sin\theta_1 - ikx\sin\theta}$ must also be independent of x. Thus,

$$k_1\sin\theta_1 = k\sin\theta \tag{51}$$

which is Snell's law. Thus, the first boundary condition leads to Snell's law. The second boundary condition leads to the equation:

$$\begin{aligned} &\frac{1}{\rho}p_o e^{ikx\sin\theta}(-ik\cos\theta + Vik\cos\theta) \\ &\quad = \frac{1}{\rho_1}Wp_o e^{ik_1 x\sin\theta_1}(-ik_1\cos\theta_1) \end{aligned} \tag{52}$$

Substituting Eq. (51) into (50) gives:

$$1 + V = W \tag{53}$$

and substituting Eq. (53) into (52) gives:

$$\begin{aligned} &\frac{1}{\rho}e^{ikx\sin\theta}(-ik\cos\theta + Vik\cos\theta) \\ &\quad = \frac{1}{\rho_1}(1 + V)e^{ik_1 x\sin\theta_1}(-ik_1\cos\theta_1) \end{aligned} \tag{54}$$

or,

$$\frac{\rho_1}{\rho}(-ik\cos\theta + Vik\cos\theta) = (1 + V)(-ik_1\cos\theta_1)$$

So

$$V = \frac{(\rho_1/\rho)k\cos\theta - k_1\cos\theta_1}{(\rho_1/\rho)k\cos\theta + k_1\cos\theta_1} \tag{55}$$

$$= \frac{\frac{\rho_1}{\rho}\cos\theta - \frac{k_1}{k}\cos\theta_1}{\frac{\rho_1}{\rho}\cos\theta + \frac{k_1}{k}\cos\theta_1} \qquad \frac{k_1}{k} = \frac{c}{c_1} \tag{56}$$

Equations (51), (53), and (56) give the unknowns θ_1, V, and W as functions of the known quantities ρ_1, ρ, c_1, c, and θ. Note that if the two media are the same then $V = 0$ and $W = 1$. Thus, there is no reflection, and the transmission is

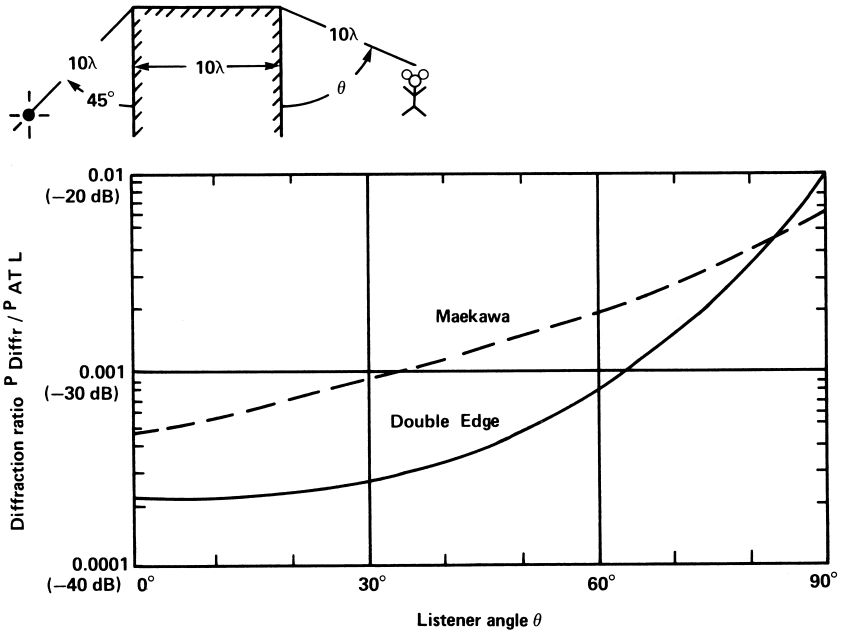


FIGURE 7 Diffraction over a wide barrier. Plot of the ratio of square of diffracted sound pressure amplitude p_{Diffr} to the square of the amplitude $p_{AT\,L}$ expected at an equivalent distance L from the source in the absence of the barrier. Source and listener locations are as indicated in the sketch with $z_s = z_L$ on opposite sides of a rectangular three-sided barrier. Computations based on the Maekawa approximation and on the double-edge diffraction theory are presented for listener angle θ between 0° and 90°. Here, L represents a distance of 30 wavelengths ($10\lambda + 10\lambda + 10\lambda$). (Reprinted by permission from Pierce, A. D. (1974). *J. Acoust. Soc. Am.* **55** (5), 953.)

100%; that is, the incident wave continues to move along the original path. As $\theta \to \pi/2$ (grazing incidence) then $V \to -1$ and $W \to 0$. This says that there is no wave transmitted to the second medium at grazing incidence of the incident wave. For θ such that $(\rho_1/\rho)\cos\theta = (k_1/k)\cos\theta$, the reflection coefficient vanishes, and there is complete transmission similar to the case in which the two media are the same.

C. Diffraction

One of the most interesting practical applications of diffraction is in barriers. Figure 7 shows results of diffraction over a wide barrier.

This plot illustrates how sound bends around corners. As the listener gets closer to the barrier (i.e., as $\theta \to 0$), the sound is reduced by almost 40 db for the case shown. When the listener is at the top of the barrier ($\theta \to 90°$), the reduction is only 20 db. In the first case ($\theta \to 0$), the sound has to bend around the two corners. However, for the second case, it has to bend only around the first corner. Such barriers are often used to block the noise from superhighways to housing developments. As can be seen from the curve, the listener has to be well in the shadow zone to achieve maximum benefits.

D. Interference

If the pressure in two different acoustic waves is p_1, p_2 and the velocity of the waves is u_1, u_2, respectively, then the intensity I for each of the waves is

$$I_1 = p_1 u_1 \qquad I_2 = p_2 u_2 \tag{57}$$

When the waves are combined, the pressure p in the combined wave is

$$p = p_1 + p_2 \tag{58}$$

and the velocity u in the combined wave is

$$u = u_1 + u_2 \tag{59}$$

The intensity of the combined wave is

$$\begin{aligned} I &= pu \\ &= (p_1 + p_2)(u_1 + u_2) = p_1 u_1 + p_2 u_2 + p_2 u_1 + p_1 u_2 \\ &= I_1 + I_2 + (p_2 u_1 + p_1 u_2) \end{aligned} \tag{60}$$

Equation (60) states that the sum of the intensities of the two waves is not merely the sum of the intensities of each of the waves, but that there is an extra term. This term is called the interference term. The phenomena that the superposition principle does not hold for intensity in linear acoustics is known as interference. If both u_1 and

u_2 are positive, then what results is called constructive interference. If $u_1 = -u_2$, then $I = 0$ and what results is called destructive interference.

E. Scattering

The discussion of reflection, refraction, and transmission was limited to waves that impinged on a flat infinite surface such as the interface between two fluids. In those cases, the phenomena of reflection, refraction, and transmission were clear cut, and the various phenomena could be separated.

If the acoustic wave encounters a finite object, the processes are not so clear cut and cannot be separated. The process of scattering actually involves reflection, refraction, and transmission combined, but it is called scattering because the wave scatters from the object in all directions.

Consider the classical two-dimensional problem of a plane wave impinging on a rigid cylinder, as shown in Fig. 8. The intensity of the scattered wave can be written:

$$I_s \approx \left(\frac{2I_o a}{\pi r}\right)|\psi_s(\theta)|^2 \tag{61}$$

where I_o is the intensity of the incident wave ($I_o = P_o^2/2\rho_o c_o$ where P_o is the pressure in the incident wave, ρ_o is the density of the medium, and c_o is the sound velocity in the medium), and $\psi_s(\theta)$ is a distribution function. Figure 9 shows the scattered power and distribution in intensity for various values of ka.

Several interesting cases can be noted. If $ka \to 0$, then the wavelength of the sound is very large compared with the radius of the cylinder, and the scattered power goes to zero. This means that the sound travels as if the object were not present at all. If the wavelength is very small compared with the cylinder radius, it can be shown that most of the scattering is in the backward direction in the form of an echo or reflection, in the same manner as would occur at normal incidence of a plane wave on an infinite plane. Thus, for small ka (low frequency), there is mostly forward scattering, and for large ka (high frequency), there is mostly backscattering.

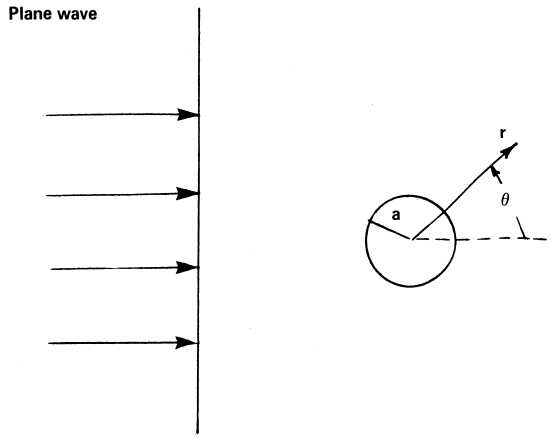


FIGURE 8 Plane wave impinging on a rigid cylinder. I_o = intensity of incident plane wave; a = radius of cylinder; r, θ = cylindrical coordinates of field point.

Consider now the contrast between scattering from elastic bodies compared with rigid bodies. Let a plane wave of magnitude p and frequency ω impinge broadside on a cylinder as shown in Fig. 10(a). Let $f_\infty(\theta)$ be defined as follows:

$$f_\infty(\theta) = \left(\frac{2r}{a}\right)^{1/2}\frac{p_s(\theta)}{p_o} = \text{form function}$$

where r = radial distance to the point where the scattered pressure is being measured; a = outside radius of the cylinder; b = inside radius of a shell whose outside radius is a; $p_s(\theta)$ = amplitude of scattered pressure; p_o = amplitude of incident wave; $ka = \omega a/c_o$; $\omega = 2\pi f$; f = frequency of incoming wave; c_o = sound velocity in the medium.

Figure 10(b) shows the form function for a rigid cylinder as a function of ka. Figure 10(c) shows this function for a rigid sphere of outside radius a. Contrast this with Fig. 10(d), which gives the form function for a solid aluminum cylinder in water, and with Fig. 10(e), which shows the function for elastic aluminum shells of various thicknesses. As one can see, the elasticity of the body has a dominant effect on the acoustic scattering from the body.

VI. FREE AND CONFINED WAVES

A. Propagating Waves

The acoustic wave equation states that the pressure satisfies the equation:

$$c^2\nabla^2 p = \frac{\partial^2 p}{\partial t^2} \tag{62}$$

For illustrative purposes, consider the one-dimensional case in which the waves are traveling in the x direction. The equation satisfied in this case is

$$c^2\frac{\partial^2 p}{\partial x^2} = \frac{\partial^2 p}{\partial t^2} \tag{63}$$

The most general solution to this equation can be written in the form:

$$p = f_1(x + ct) + f_2(x - ct) \tag{64}$$

This solution consists of two free traveling waves moving in opposite directions.

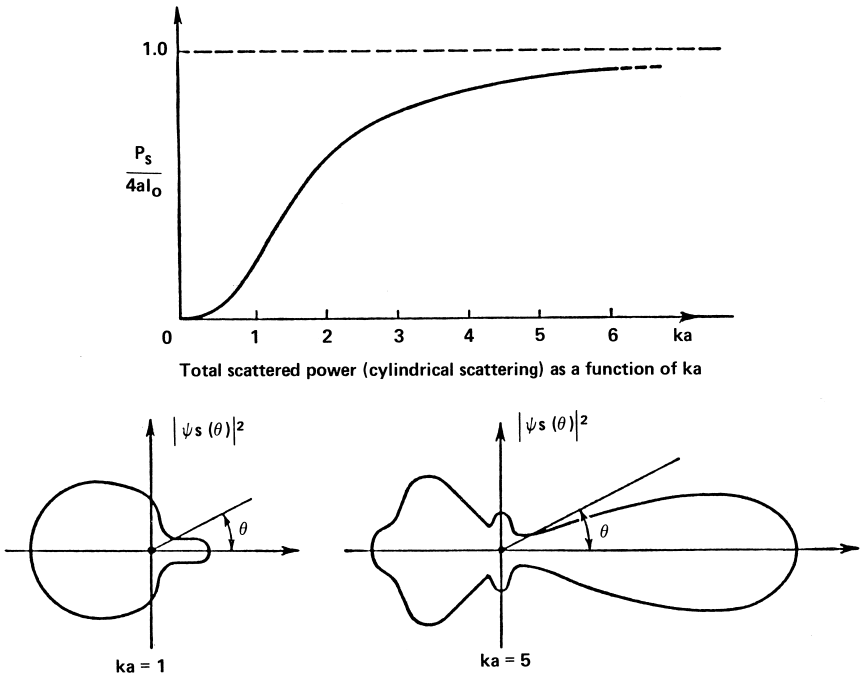


FIGURE 9 Scattered power and distribution in intensity for a rigid cylinder. $\psi_S(\theta) =$ angular distribution function; $a =$ radius of cylinder, $I_o =$ incident intensity of plane wave; $k = \omega/c_o$; $\omega =$ frequency of wave; $c_o =$ sound velocity in the medium. (Reprinted by permission from Lindsay, R. B. (1960). "Mechanical Radiation," McGraw-Hill, New York.)

B. Standing Waves

Consider the waves described immediately above, but limit the discussion to harmonic motion of frequency ω. One of the waves takes the form:

$$p = A\cos(kx - \omega t) \tag{65}$$

where $k = \omega/c = 2\pi/\lambda$ and $\lambda =$ wavelength of the sound. This equation can also be written in the form:

$$p = A\cos k(x - ct) \tag{66}$$

If this wave hits a rigid barrier, another wave of equal magnitude is reflected back toward the source of the wave motion. The reflected wave is of the form:

$$p = A\cos k(x + ct) \tag{67}$$

If the source continues to emit waves of the form of Eq. (66), and reflections of the form of Eq. (67) come back, then the resulting pattern is a superposition of the waves, that is,

$$p = A\cos(kx + \omega t) + A\cos(kx - \omega t) \tag{68}$$

or

$$p = 2A\cos kx\cos\omega t$$

The resultant wave pattern no longer has the characteristics of a traveling wave. The pattern is stationary and is known as a standing wave.

C. Reverberation

When traveling waves are sent out in an enclosed space, they reflect from the walls and form a standing wave pattern in the space. This is a very simple description of a very complicated process in which waves impinge on the walls from various angles, reflect, and impinge again on another wall, and so on. The process of reflection takes place continually, and the sound is built up into a sound field known as a reverberant field. If the source of the sound is cut off, the field decays. The amount of time that it takes for the sound energy density to decay by a factor of 10^6 (i.e., 60 db) is called the reverberation time of the room. The sound energy density concept was used by Sabine in his fundamental discoveries on reverberation. He found that sound fills a reverberant room in such a way that the average energy per unit volume (i.e., the energy density) in any region is nearly the same as in any other region.

The amount of reverberation depends on how much sound is absorbed by the walls in each reflection and in the air. The study of room reverberation and the answering of questions such as how much the sound is absorbed by people in the room, and other absorbers placed in the room, are included in the field of architectural acoustics. The acoustical design of concert halls or any structures in which sound and reverberation are of importance is a specialized and intricate art.

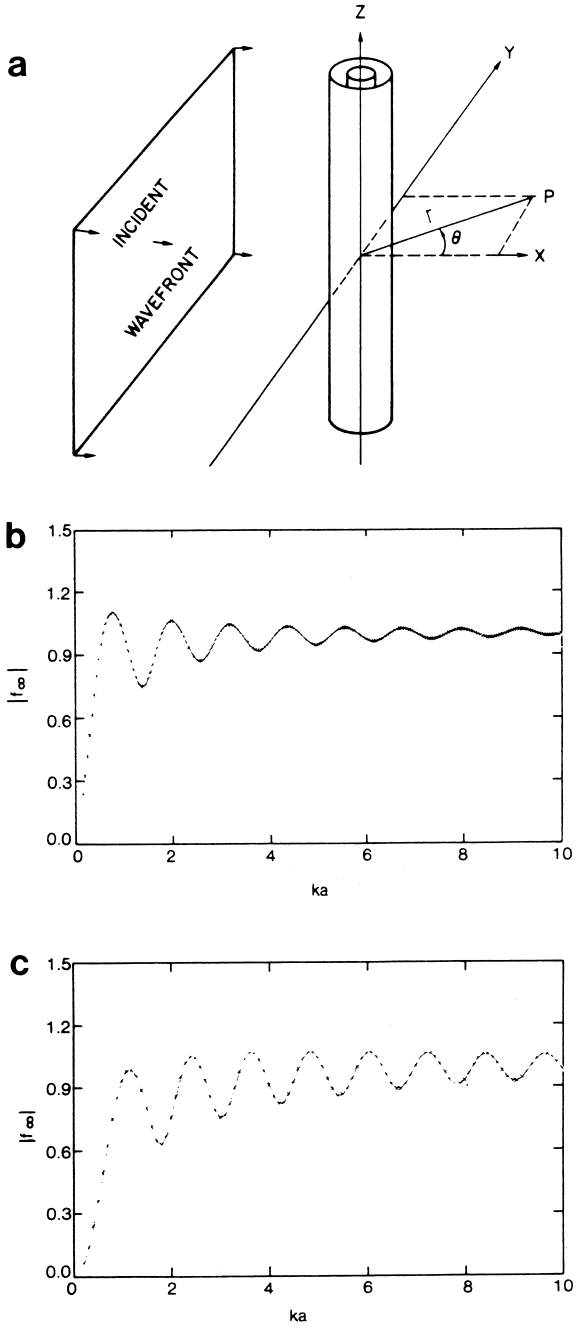


FIGURE 10 (a) The geometry used in the description of the scattering of a plane wave by an infinitely long cylinder. (b) The form function for a rigid cylinder. (c) The form function vs. *ka* for a rigid sphere.

D. Wave Guides and Ducts

When a wave is confined in a pipe or duct, the duct is known as a wave guide because it prevents the wave from moving freely into the surrounding medium. In discussing normal mode theory in connection with underwater sound propagation, the boundary conditions were stipulated on the surface and bottom. The problem thus became one of propagation in a wave guide.

One wave guide application that leads to interesting implications when coupled and uncoupled systems are considered, is the propagation of axially symmetric waves in a fluid-filled elastic pipe. If axially symmetric pressure waves of magnitude p_o and frequency ω are sent out from a plane wave source in a fluid-filled circular pipe, the pressure at any time t and at any location x from the source can be written as follows:

$$p \approx p_o\left[1 + i\omega\left(\frac{r^2}{2ac\zeta_t}\right)\right]e^{-(x\kappa_t/a)+i[\omega/c+(\sigma_t/a)]x-i\omega t} \tag{69}$$

where r is the radial distance from the center of the pipe to any point in the fluid, a the mean radius of the pipe, c the sound velocity in the fluid inside the pipe, ω the radian frequency of the sound, x the longitudinal distance from the disturbance to any point in the fluid, and z_t the impedance of the pipe such that:

$$\frac{1}{z_t} = \frac{1}{\rho c\zeta_t} = \frac{1}{\rho c}(\kappa_t - i\sigma_t) \tag{70}$$

where $\kappa_t/\rho c$ is the conductance of the pipe and $\sigma_t/\rho c$ the susceptance of the pipe. The approximate value of the wave velocity υ down the tube is

$$\upsilon = c[1 - \sigma_t(\lambda/2\pi a)] \tag{71}$$

If the tube wall were perfectly rigid, then the tube impedance would be infinite ($\sigma_t \to 0$) and the velocity would be c. The attenuation is given by the factor $e^{i\kappa_t x/a}$. If the tube were perfectly rigid ($\kappa_t = 0$), then the attenuation would be zero. If the tube is flexible, then energy gradually leaks out as the wave travels and the wave in the fluid attenuates. This phenomenon is used extensively in trying to reduce sound in tubes and ducts by using acoustic liners. These acoustic liners are flexible and absorb energy as the wave travels down the tube.

One critical item must be mentioned at this point. It has been assumed that the tube impedance (or conductance and susceptance) can be calculated independently. This is an assumption that can lead to gross errors in certain cases. It will be discussed further when coupled systems are considered.

The equation for an axisymmetric wave propagating in a rigid circular duct or pipe is as follows:

$$p(r, z) = p_{m0}J_0(\alpha_{om}r/a)e^{i(\gamma_{om}z-\omega t)}$$

p_{mo} is the amplitude of the pressure wave, r is the radial distance from the center of the pipe to any point, a is the radius of the pipe, z is the distance along the pipe, ω is the radian frequency, and t is time.

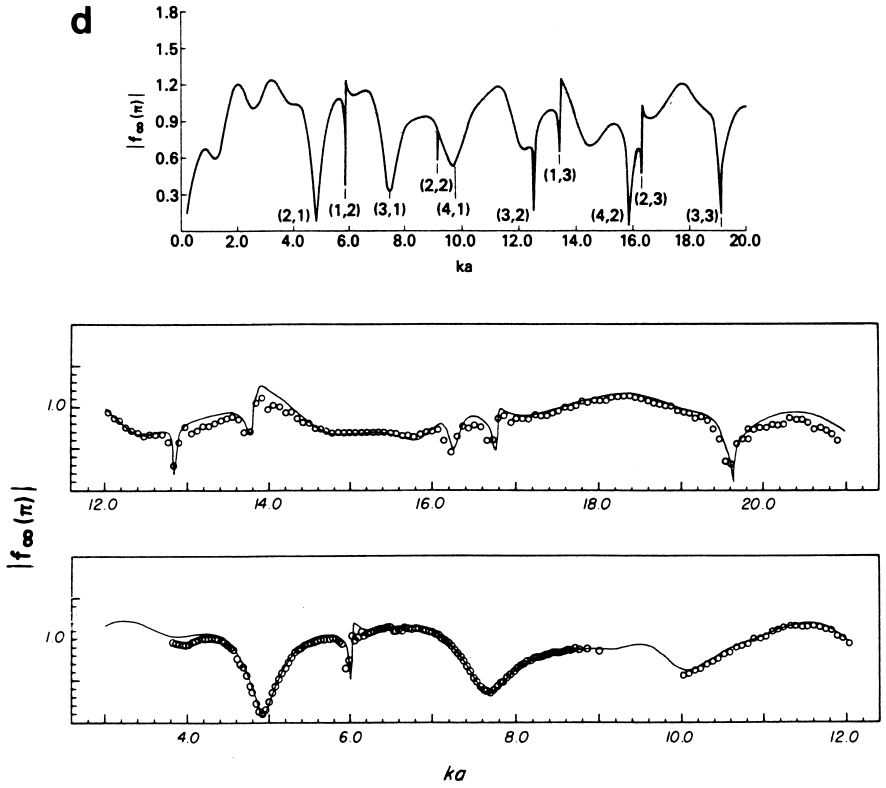


FIGURE 10 (*continued*). (d) Top, the form function for an aluminum cylinder in water, Bottom, comparison of theory (——) and experimental observation (the points) for an aluminum cylinder in water.

γ_{om} and α_{om} are related by the following formula:

$$\gamma_{om} = \left[k^2 - (\alpha_{om}/a)^2\right]^{1/2}$$

J_0 is the Bessel Function of order 0.

$$k = 2\pi f/c_i$$

In the above relation, f is the frequency of the wave and c_i is the sound velocity in the fluid inside the pipe (for water this sound velocity is about 5000 ft/sec and for air it is about 1100 ft/ sec). The values of α_{om} for the first few m are

$$m = 0 \qquad \alpha_{00} = 0$$
$$m = 1 \qquad \alpha_{01} = 3.83$$
$$m = 2 \qquad \alpha_{02} = 7.02$$

If $k < \alpha_{om}/a$ then γ_{om} is a pure imaginary number and the pressure takes the form:

$$p(r, z) = P_{m0} J_0(\alpha_{om} r/a) e^{-\gamma} om^z e^{-i\omega t}$$

which is the equation for a decaying wave in the z direction. For frequencies which give $k < \alpha_{om}/a$, no wave is propagated down the tube. Propagation takes place only for frequencies in which $k > \alpha_{om}/a$. Since $\alpha_{00} = 0$, propagation always takes place for this mode. The frequency at which γ_{om} is 0 is called the cutoff frequency and is as follows:

$$f_{om} = c_i \alpha_{om}/2\pi a$$

For frequencies below the cutoff frequency, no propagation takes place. For general asymmetric waves the pressure takes the form:

$$p(r, z, \phi) = P_{nm} J_n(\alpha_{nm} r/a) e^{i(\gamma_{nm} z - \omega t)} \cos n\phi$$

where $\gamma_{nm} = [k^2 - (\alpha_{nm}/a)^2]^{1/2}$, $f_{nm} = c_i \alpha_{nm}/2\pi a$.

A few values for α_{nm} for $n > 0$ are as follows:

$$\alpha_{10} = 1.84 \qquad \alpha_{11} = 5.31 \qquad \alpha_{12} = 8.53$$
$$\alpha_{20} = 3.05 \qquad \alpha_{21} = 6.71 \qquad \alpha_{22} = 9.97$$

It is seen that only $\alpha_{00} = 0$, and this is the only mode that propagates at all frequencies regardless of the size of the duct.

Consider a 10-in.-diameter pipe containing water. The lowest cutoff frequency greater than the 00 mode is

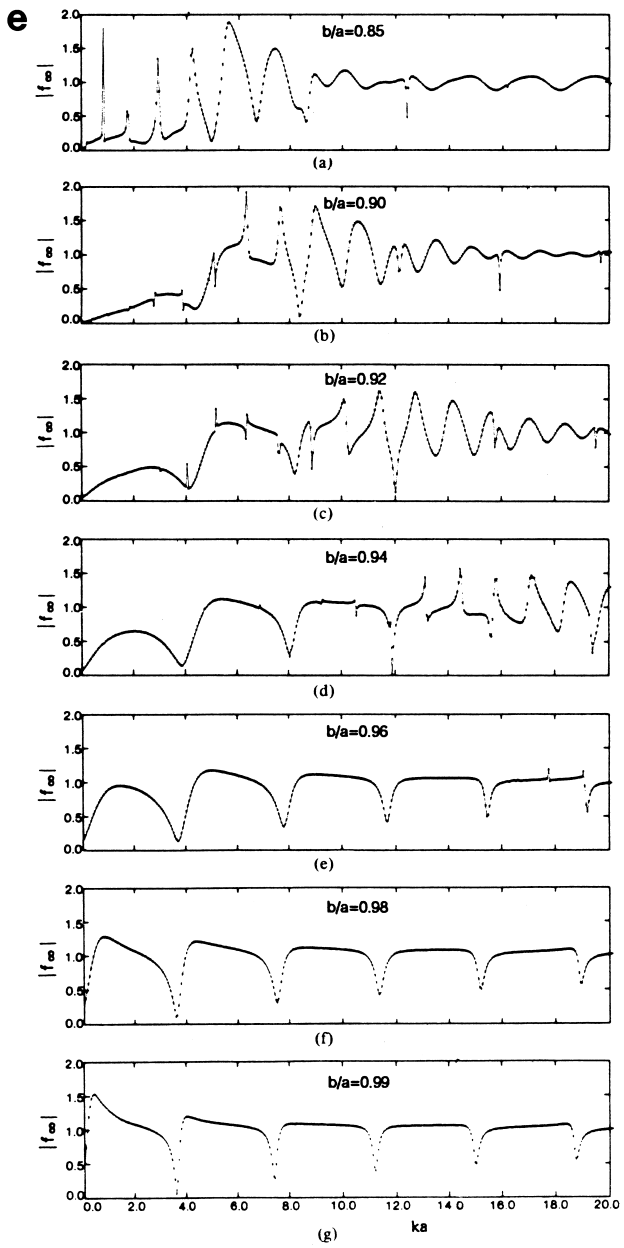


FIGURE 10 (*continued*). (e) The form function vs. *ka* over the range of $0.2 \leq ka \leq 20$ for aluminum shells with *b*/*a* values of (a) 0.85, (b) 0.90, (c) 0.92, (d) 0.94, (e) 0.96, (f) 0.98, and (g) 0.99. (Figs. 10(a–e) reprinted by permission from Neubauer, W. G., (June 1986). "Acoustic Reflection from Surfaces and Shapes," Chapter 4, Eq. (6) and Figs. 1, 2, 7(a), 13, and 27, Naval Research Laboratory, Washington, D.C.)

3516 Hz. Thus, nothing will propagate below 3516 Hz except the lowest axisymmetric mode (i.e., the 00 mode). If the pipe were 2 in. in diameter then nothing would propagate in the pipe below 17,580 Hz except the 00 mode. This means that in a great many practical cases no matter what is exciting the sound inside the duct, only the lowest axisymmetric mode (00 mode) will propagate in the duct.

VII. SOUND RADIATION AND VIBRATION

A. Helmholtz Integral, Sommerfeld Radiation Condition, and Green's Functions

In this presentation, acoustic fields that satisfy the wave equation of linear acoustics are of interest.

$$c^2\nabla^2 p = \partial^2 p/\partial t^2 \tag{72}$$

where p is the pressure in the field at point P and at time t. For sinusoidal time-varying fields,

$$p(P,t) = p(P)e^{i\omega t} \tag{73}$$

so that p satisfies the Helmholtz equation:

$$\nabla^2 p + k^2 p = 0 \qquad k^2 = \omega^2/c^2 \tag{74}$$

This Helmholtz equation can be solved in general form by introducing an auxiliary function called the Green's function. First, let φ and ψ be any two functions of space variables that have first and second derivatives on S and outside S (see Fig. 11). Let V_o be the volume between S_o and the boundary at ∞. Green's theorem states that within the volume V_o,

$$\int_S \left(\varphi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\varphi}{\partial n}\right) dS = \int_{V_o} (\varphi\nabla^2\psi - \psi\nabla^2\varphi)\, dV_o \tag{75}$$

where S denotes the entire boundary surface and V_o the entire volume of the region outside S_o. In Eq. (75) $\partial/\partial n$ denotes the normal derivative at the boundary surface. Rearrange the terms in Eq. (75) and subtract $\int_{V_o} k^2\varphi\psi\, dV_o$ from each side, and the result is

$$\int_S \varphi\frac{\partial\psi}{\partial n}\, dS - \int_{V_o} \varphi(\nabla^2\psi + k^2\psi)\, dV_o = \int_S \psi\frac{\partial\varphi}{\partial n}\, dS - \int_{V_o} \psi(\nabla^2\varphi + k^2\varphi)\, dV_o \tag{76}$$

Now choose φ as the pressure p in the region V_o; thus,

$$\nabla^2 p + k^2 p = \nabla^2\varphi + k^2\varphi = 0 \tag{77}$$

and choose ψ as a function that satisfies:

$$\nabla^2\psi + k^2\psi = \delta(P - P') \tag{78}$$

where $\delta(P - P')$ is a δ function of field points P and P'. Choose another symbol for ψ, that is,

$$\psi = g(P, P', \omega) \tag{79}$$

By virtue of the definition of the δ function, the following is obtained:

$$\int_{V_o} \varphi(P')\delta(P - P')\, dP' = \varphi(P) \tag{80}$$

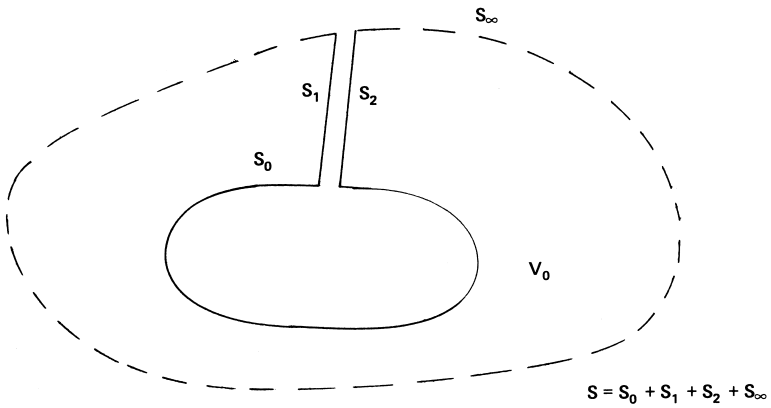


FIGURE 11 The volume and boundary surface. S_∞ = surface at ∞; S_o = radiating surface; V_0 = volume between S_∞ and S_o; S_1, S_2 = surfaces connecting S_o to S_∞.

Thus, Eq. (76) becomes:

$$\int_S p(S,\omega)\frac{\partial g(P,S,\omega)}{\partial n} - p(P,\omega) = \int_S g(P,S,\omega)\frac{\partial p(S,\omega)}{\partial n}\,dS \tag{81}$$

or

$$p(P,\omega) = \int_S \left[p(S,\omega)\frac{\partial g(P,S,\omega)}{\partial n} - g(P,S,\omega)\frac{\partial p(S,\omega)}{\partial n}\right] dS \tag{82}$$

It is now clear that the arbitrary function ψ was chosen so that the volume integral would reduce to the pressure at P. The function g is the Green's function, which thus far is a completely arbitrary solution of:

$$\nabla^2 g(P,P',\omega) + k^2 g(P,P',\omega) = \delta(P-P') \tag{83}$$

For this Green's function to be a possible solution, it must satisfy the condition that there are only outgoing traveling waves from the surface S_o to ∞, and no waves are coming in from ∞. Sommerfeld formulated this condition as follows:

$$\lim_{r\to\infty} r\left(\frac{\partial g}{\partial r} - ikg\right) = 0 \tag{84}$$

where r is the distance from any point on the surface to any point in the field. A solution that satisfies Eqs. (83) and (84) can be written as follows:

$$g = \frac{1}{4\pi}\frac{e^{ikr}}{r} \tag{85}$$

This function is known as the free-space Green's function. Thus, Eq. (82) can be written in terms of this free-space Green's function as follows:

$$p(P,\omega) = \frac{1}{4\pi}\int_S \left[\rho(S,\omega)\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial p(S,\omega)}{\partial n}\right] dS \tag{86}$$

Several useful alternative forms of Eqs. (82) and (86) can be derived. If a Green's function can be found whose normal derivative vanishes on the boundary S_o, then:

$$\frac{\partial g}{\partial n}(P,S,\omega) = 0 \qquad \text{on } S_o$$

and Eq. (82) becomes:

$$p(P,\omega) = -\int_S g(P,S,\omega)\frac{\partial p}{\partial n}\,dS \tag{87}$$

Alternatively, if a Green's function can be found that itself vanishes on the boundary S_o, then $g(P,S,\omega) = 0$ on S_o and Eq. (82) becomes:

$$p(P,\omega) = \int_S p(S,\omega)\frac{\partial g(P,S,\omega)}{\partial n}\,dS \tag{88}$$

From Newton's law,

$$\partial p/\partial n = -\rho\ddot{w}_n \tag{89}$$

where $\ddot{w}_n$ is the normal acceleration of the surface. Thus, Eq. (87) can be written as:

$$p(P,\omega) = \rho\int_S g(P,S,\omega)\ddot{w}_n\,dS \tag{90}$$

If γ is the angle between the outward normal to the surface at S and the line drawn from S to P, then Eq. (86) can be written in the form (assuming harmonic motion):

$$p(P,\omega) = \frac{1}{4\pi}\int_S (\rho\ddot{w}_n(s) - ikp(s)\cos\gamma)\frac{e^{ikr}}{r}\,dS \tag{91}$$

Since $p(s)$ is unknown, then Eq. (86) or, alternatively, Eq. (91) is an integral equation for p.

An interesting limiting case can be studied. Assume a low-frequency oscillation of the body enclosed by S_o. For this case, k is very small and Eq. (91) reduces to:

$$p(P,\omega) = \frac{1}{4\pi r}\int_S \rho\ddot{w}_n\, ds \tag{92}$$

or

$$p(P,\omega) = \rho\ddot{V}/4\pi R \tag{93}$$

where $\ddot{V}$ is the volume acceleration of the body.

For some cases of slender bodies vibrating at low frequency, it can be argued that the term involving p in Eq. (91) is small compared with the acceleration term. For these cases,

$$p(P,\omega) \approx \frac{1}{4\pi}\int_S \rho\ddot{w}_n(S)\frac{e^{ikr}}{r}\, dS \tag{94}$$

B. Rayleigh's Formula for Planar Sources

It was stated in the last section that if a Green's function could be found whose normal derivative vanishes at the boundary, then the sound pressure radiated from the surface could be written as Eq. (87) or, alternatively, using Eq. (89):

$$p(P,\omega) = \rho\int_S g(P,S,\omega)\ddot{w}_n(S)\, ds \tag{95}$$

It can be shown that such a Green's function for an infinite plane is exactly twice the free-space Green's function, that is,

$$g(P,S,\omega) = 2\times\frac{1}{4\pi}\frac{e^{ikr}}{r} \tag{96}$$

The argument here is that the pressure field generated by two identical point sources in free space displays a zero derivative in the direction normal to the plane of symmetry of the two sources.

Substituting Eq. (96) into Eq. (95) gives:

$$p = \frac{\rho}{2\pi}\int_S \frac{e^{ikr}}{r}\ddot{w}_n(S)\, ds \tag{97}$$

Equation (97) is known as Rayleigh's formula for planar sources.

C. Vibrating Structures and Radiation: Multipole Contributions

To make it clear how the above relations can be applied, consider the case of a slender vibrating cylindrical surface

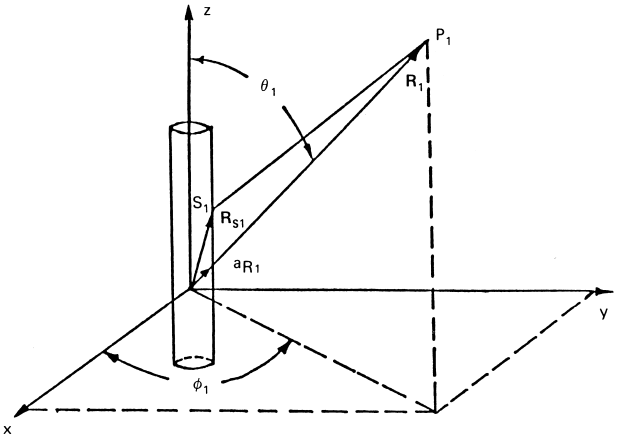


FIGURE 12 Geometry of the vibrating cylinder. $\mathbf{R}_{S_1}$ = radius vector to point S_1 on the cylinder surface; $\mathbf{R}_1$ = radius vector to the far-field point; $\mathbf{a}_{R_1}$ = unit vector in the direction of $\mathbf{R}_1$; P_1 = far-field point (with spherical coordinates R_1, φ_1, θ_1). (Reprinted by permission from Greenspon, J. E. (1967). *J. Acoust. Soc. Am.* **41** (5), 1203.)

at low frequency such that Eq. (94) applies. The geometry of the problem is shown in Fig. 12.

$$\frac{e^{ikr}}{r} = \frac{e^{ikR_1}}{R_1}e^{-ik(\mathbf{a}_{R_1}\cdot\mathbf{R}_{S_1})}$$
$$\mathbf{a}_{R_1}\cdot\mathbf{R}_{S_1} = z_o\cos\theta_1 + x_o\sin\theta_1\cos\varphi_1 + y_o\sin\theta_1\sin\varphi_1 \tag{98}$$

where x_o, y_o, z_o are the rectangular coordinates of a point on the vibrating surface of the structure, $\mathbf{R}_{S_1}$ is the radius vector to point S_1 on the surface, R_1, θ, φ_1 are the spherical coordinates of point P_1 in the far field, any $\mathbf{a}_{R_1}$ is a unit vector in the direction of $\mathbf{R}_1$ (the radius vector from the origin to the far field point). Thus, $\mathbf{a}_{R_1}\cdot\mathbf{R}_{S_1}$ is the projection of $\mathbf{R}_{S_1}$ on $\mathbf{R}_1$ making $R_1 - \mathbf{a}_{R_1}\cdot\mathbf{R}_{S_1}$ the distance from the far field point to the surface point.

Assume the acceleration distribution of the cylindrical surface to be that of a freely supported shell subdivided into its modes, that is,

$$\ddot{w}_2(S) = \sum_{m=1}^{\infty}\sum_{q=0}^{\infty}(A_{mq}\cos q\varphi_1 + B_{mq}\sin q\varphi)\sin\frac{m\pi z}{l} \tag{99}$$

Expression (99) is a half-range Fourier expansion in the longitudinal z direction (which is chosen as the distance along the generator) between bulkheads, which are distance 1 apart. The expression (99) is a full-range Fourier expansion in the peripheral φ direction. It is known that such an expression does approximately satisfy the differential equations of shell vibration and practical end conditions at the edges of the compartment. Substitution of Eqs. (98) and (99) into (94) and integrating results in:

$$p(P_1,\omega)=\frac{\rho_0 e^{ikR_1}}{4\pi R_1}\sum_{m=1}^{\infty}\sum_{q=0}^{\infty}\{A_{mq}\cos q\varphi_1+B_{mq}\sin q\varphi_1\}$$

$$\times J_q(ka\sin\theta_1)2\pi a l i^q\left\{\left[\frac{1}{2(m\pi-kl\cos\theta_1)}\right.\right.$$

$$+\frac{1}{2(m\pi+kl\cos\theta_1)}\Bigg]-\left[\frac{\cos(m\pi-kl\cos\theta_1)}{2(m\pi+kl\cos\theta_1)}\right.$$

$$+\frac{\cos(m\pi-kl\cos\theta_1)}{2(m\pi+kl\cos\theta_1)}\Bigg]-i\left[\frac{\sin(m\pi-kl\cos\vartheta_1)}{2(m\pi+kl\cos\theta_1)}\right.$$

$$\left.\left.-\frac{\sin(m\pi-kl\cos\vartheta_1)}{2(m\pi+kl\cos\theta_1)}\Bigg]\right\}$$

Consider the directivity pattern in the horizontal plane of the cylindrical structure, that is, at $\varphi_1=\pi/2$, and let us examine the A_{mq} term in the series above. After some algebraic manipulation, the amplitude of the pressure can be written as follows:
For $m\pi\neq kl\cos\theta_1$,

$$I_{mq}=\frac{p_{mq}/2\pi al}{\frac{A_{mq}\rho_0 e^{ikR_1}}{4\pi R_1}\cos q\theta_1}$$

$$=\frac{\sqrt{2}m\pi J_q(ka\sin\theta_1)\sqrt{1-(-1)^m\cos(kl\cos\theta_1)}}{(m\pi)^2-(kl\cos\theta_1)^2}$$

For $m\pi=kl\cos\theta_1$,

$$I_{mq}=\tfrac{1}{2}J_q(ka\sin\theta_1) \tag{100}$$

where J_q is the Bessel function of order q.

Figure 13 shows the patterns of the far-field pressure for various values of ka, m, q. A source pattern is defined as one that is uniform in all directions. A dipole pattern has two lobes, a quadrupole pattern has four lobes, and so on. Note that for $ka=0.1$, $q=1$, $m=1, 3, 5$; all show dipole-type radiation. In general. Fig. 13 shows how the multipole contributions depend upon the spatial pattern of the acceleration of the structure.

For low frequencies where $kl\ll m\pi$, it is seen that (noting that $\pi l/\lambda=kl/2$) for m even,

$$I_{mq}\approx\frac{2J_q(ka\sin\theta_1)\sin\left(\frac{\pi l}{\lambda}\cos\theta_1\right)}{m\pi} \tag{101}$$

for m odd,

$$I_{mq}\approx\frac{2J_q(ka\sin\theta_1)\cos\left(\frac{\pi l}{\lambda}\cos\theta_1\right)}{m\pi}$$

Thus, at low frequencies (i.e., for $kl\ll m\pi$), the structural modes radiate as though there were two sources at the edges of the compartment (i.e., l apart). What results is the directivity pattern for two point sources at the edges of the compartment modified by $J_q(ka\sin\theta_1)$. If the longitudinal mode number m is even, then the sources are 180 degrees out of phase, and if m is odd, the sources are in phase. Such modes are called edge modes.

D. Vibrations of Flat Plates in Water

Consider a simply supported flat rectangular elastic plate placed in an infinite plane baffle and excited by a point force as shown in Fig. 14. The plate is made of aluminum and is square with each side being 13.8 in. long. Its thickness is $\frac{1}{4}$ in., and it has a damping loss factor of 0.05. The force is equal to 1 lb, has a frequency of 3000 cps, and is located at $x_0=9$ in., $y_0=7$ in. If the plate is stopped at the instant when the deflection is a maximum, the velocity pattern would be as shown in Fig. 14. Since the velocity is just the frequency multiplied by the deflection, this is an instantaneous picture of the plate.

VIII. COUPLING OF STRUCTURE/MEDIUM (INTERACTIONS)

A. Coupled Versus Uncoupled Systems

When a problem involving the interaction between two media can be solved without involving the solution in both media simultaneously, the problem is said to be uncoupled. One example of a coupled system is a vibrating structure submerged in a fluid. Usually the amplitude of vibration depends on the dynamic fluid pressure, and the dynamic pressure in the fluid depends on the amplitude of vibration. In certain limiting cases, the pressure on the structure can be written as an explicit function of the velocity of the structure. In these cases, the system is said to be uncoupled.

Another example is a pipe containing an acoustic liner. Sometimes it is possible to represent the effect of the liner by an acoustic impedance, as described in a previous section of this presentation. Such a theory was offered by Morse. As Scott indicated, implicit in Morse's theory is the assumption that the motion of the surface between the liner and the fluid depends only on the acoustic impedance and the local acoustic pressure, and not on the acoustic pressure elsewhere. This is associated with the concept of "local" and "extended" reaction. In a truly coupled system, the motion of the surface depends on the distribution of acoustic pressure, and, conversely, the acoustic pressure depends on the distribution of motion. Thus, the reaction of the surface and the pressure produced are interrelated at all points. The motion of the surface at point A is not only governed by the pressure at point A. There is motion at A due to pressure at B and, conversely, motion at B due

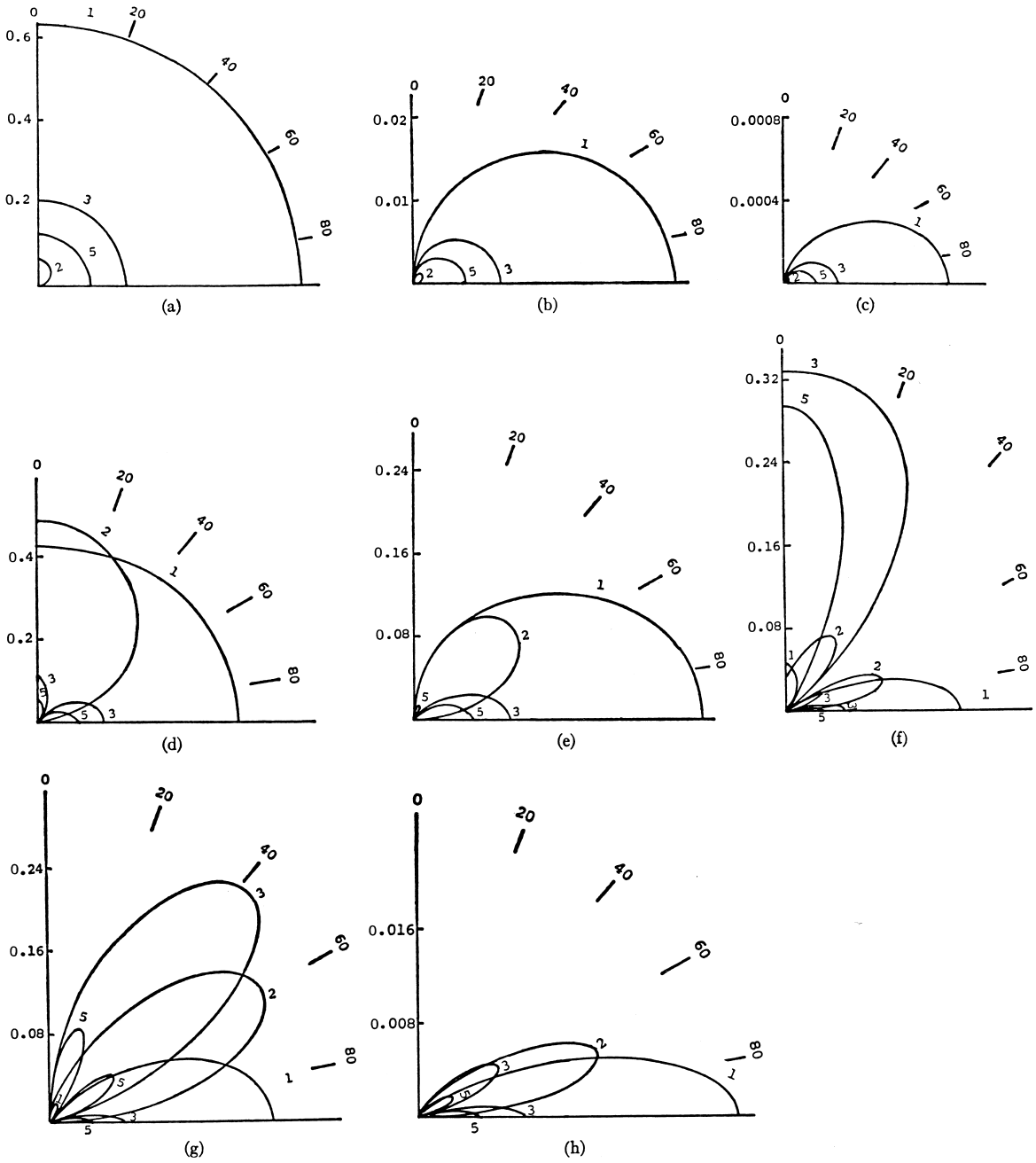


FIGURE 13 Horizontal directivity patterns for a cylinder in which $L/a = 4$, where L is length of cylinder and a is radius of cylinder. The plots show I_{mq} as a function of θ_1 at $\varphi_1 = \pi/2$ for various values of ka, m, and q. (a) $ka = 0.1$, $q = 0$; (b) $ka = 0.1$, $q = 1$; (c) $ka = 0.1$, $q = 2$; (d) $ka = 1.0$, $q = 0$; (e) $ka = 1.0$, $q = 1$; (f) $ka = 3.0$, $q = 0$; (g) $ka = 3.0$, $q = 1$; (h) $ka = 3.0$, $q = 5$. The numbers shown on the figure are the values of m. (Reprinted by permission from Greenspon, J. E. (1967). *J. Acoust. Soc. Am.* **41** (5), 1205.)

to pressure at A. Figure 15 illustrates how this assumption can lead to errors in the phase velocity and attenuation in lined ducts.

The alternative is to solve the completely coupled problem of the liner and fluid as outlined by Scott.

In aeroelastic or hydroelastic problems, it is necessary to solve the coupled problem of structure and fluid because the stability of the system usually depends on the feeding of energy from the fluid to the structure. Similarly, in acoustoelastic problems such as soft duct liners

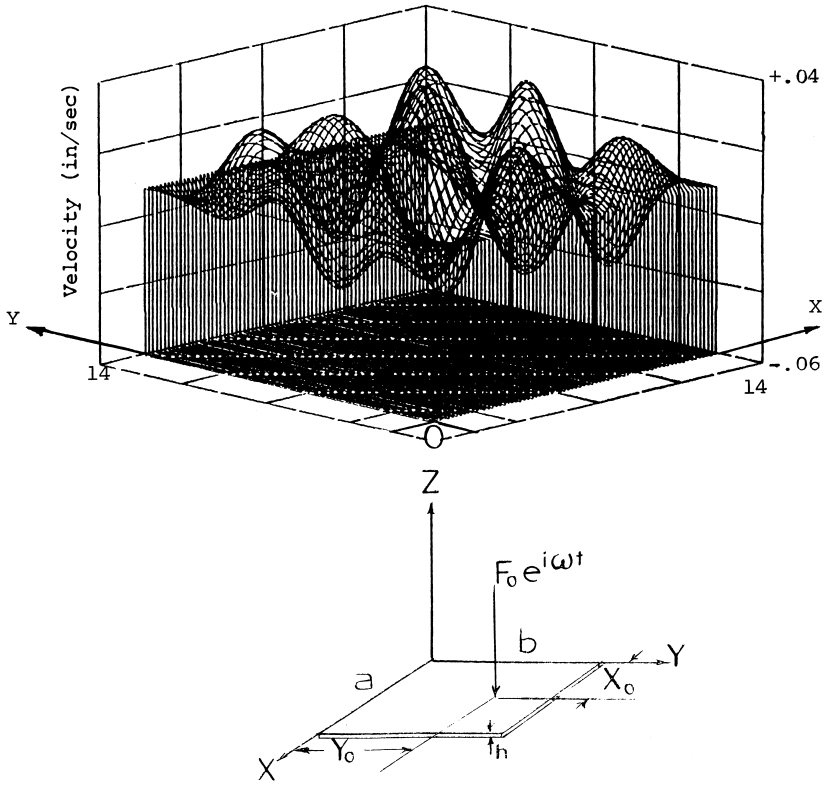


FIGURE 14 Real part of velocity of plate.

in pipes, the attenuation of the acoustic wave in the pipe is dependent upon the coupling of this acoustic wave with the wave in the liner. Similar problems were encountered with viscoelastic liners in water-filled pipes. A recent solution of the coupled problem gave realistic results, whereas the uncoupled problem gave attenuations that were much too high in much the same manner as that shown in Fig. 15.

B. Methods for Simplifying Coupled Systems

1. Impedance Considerations

In cases of acoustic waves propagating in pipes or ducts that are not quite rigid but have some elastic deformation as the wave passes, it is satisfactory to use an acoustic impedance to represent the effect of the pipe on the acoustic wave. Only when the acoustic liner or pipe is rather soft and undergoes considerable deformation during the passage of the acoustic wave is it necessary to solve the coupled problem.

It is interesting to contrast a typical uncoupled problem with a typical coupled problem. Consider a plane acoustic wave incident on a plane surface, where ζ is the dimensionless specific impedance of the surface (Fig. 16). The impedance is given by:

$$\frac{p}{\partial p/\partial n} = \frac{Z_n}{ik\rho c} = \frac{\zeta}{ik} \qquad k = w/c \tag{102}$$

In the above equation, p is the pressure, $\partial p/\partial n$ is the normal derivative of the pressure, Z_n is the normal impedance of the surface, ρ is the density of the medium, and c is the sound velocity in the medium. If p_i is the magnitude of the incident pressure and p_r is the magnitude of the reflected pressure from the surface, then the reflection coefficient R can be written as:

$$R = \frac{p_r}{p_i} = \frac{\zeta \cos\varphi_i - 1}{\zeta \cos\varphi_i + 1} \tag{103}$$

Contrast this result with the coupled problem of the reflection coefficient of an incident wave on an elastic boundary (Fig. 17). In this case, the reflection coefficient, R, and the transmission coefficient, S, are

$$R = \frac{Z_- + Z_m - Z_+}{Z_- + Z_m + Z_+} \tag{104}$$

$$S = \frac{2Z_-}{Z_+ + Z_m + Z_-} \tag{104a}$$

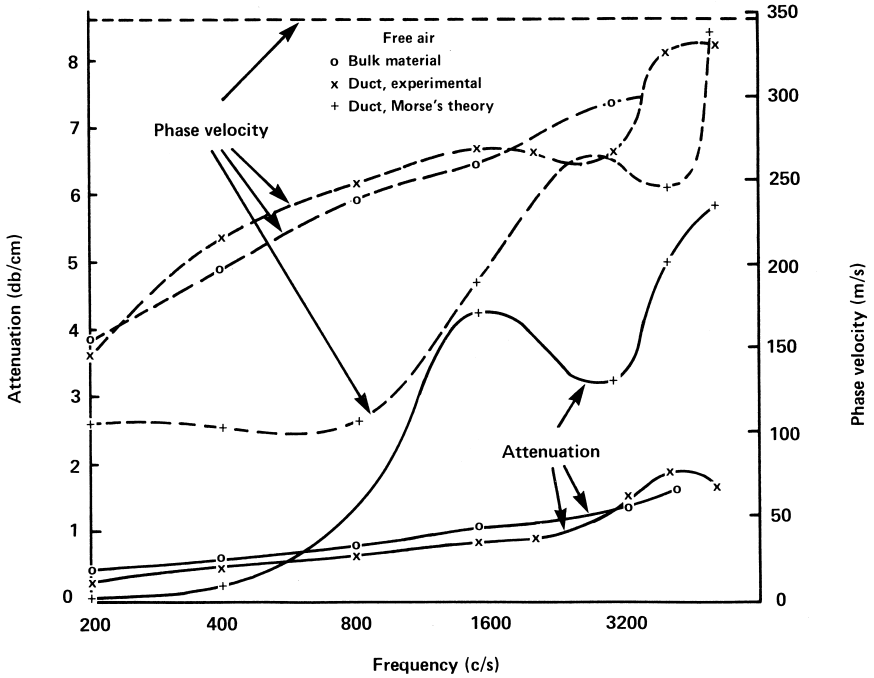


FIGURE 15 Comparison of the results of experiment and Morse's theory for a narrow lined duct. (Reprinted by permission from Scott, R. A. (1946). *Proc. Phys. Soc.* **58,** 358.)

where

$$Z_- = \rho_- c_- / \cos\varphi_t$$

$$Z_+ = \rho_+ c_+ / \cos\varphi_i$$

$$\cos\varphi_t = \sqrt{1 - \left(\frac{c_-}{c_+}\right)\sin^2\varphi_i}$$

$$Z_m = i\omega\delta\left[\left(\frac{c_m}{c_+}\right)^2 \sin^4\varphi_i - 1\right]$$

if the elastic surface is a membrane

$$Z_m = Z_p = i\omega\delta\left[\left(\frac{c_p}{c_+}\right)^4 \sin^4\varphi_i - 1\right]$$

if the elastic surface is a plate.

Where δ is the mass per unit area of surface, ρ_- the density of the lower medium, ρ_+ the density of the upper medium, c_- the sound velocity in the lower medium, c_+ the sound velocity in the upper medium, $c_m = \sqrt{T/\delta}$ the sound velocity in the membrane, T the tension in the membrane, $c_\rho = c_+\sqrt{\omega/\nu_+}$ the sound velocity in the plate, $\nu_+^2 = 12\delta(12 - \bar{\nu}^2)c_+^4/Eh^3$, $\bar{\nu}$ is Poisson's ratio for the plate material, and h the thickness of the plate.

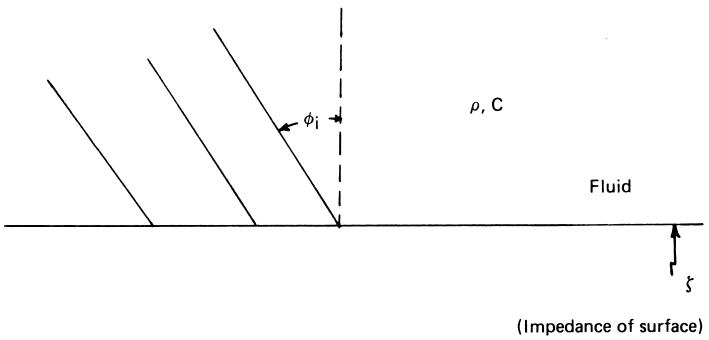


FIGURE 16 Plane wave incident on impedance surface. ζ = impedance of surface; ρ, c = density and sound velocity in medium; φ_i = angle of incidence.

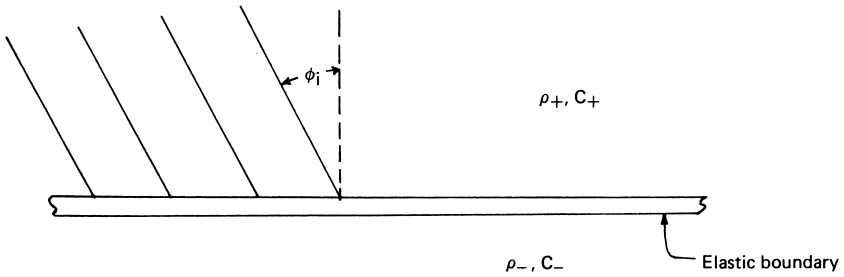


FIGURE 17 Plane wave incident on elastic boundary. $\rho_+, c_+ =$ density and sound velocity in the upper medium; $\rho_-, c_- =$ density and sound velocity in the lower medium; $\varphi_i =$ angle of incidence.

It is seen that the reflection coefficient for the coupled problem depends on both the media and the characteristics of the surface in between the media. In the uncoupled problem, the reflection coefficient depended only on the impedance of the surface.

Figure 18 illustrates the reflection and transmission coefficients between 0 and 15,000 Hz for $\frac{3}{8}$-in., $\frac{1}{2}$-in., and 1-in. steel plates with water on both sides and with water on one side and air on the other.

Note that, with water on both sides, most of the sound gets transmitted through the plate at low frequency (below 2000 Hz), whereas most of the sound gets reflected at 15,000 Hz. For plates with air on one side and water on the incoming wave side, the plate acts like a perfect baffle—i.e., all energy gets reflected back into the water and no transmission to the air takes place.

2. Asymptotic Approximations

In classical scattering problems, solutions to the Helmholtz equation are sought,

$$\nabla^2 p + k^2 p = 0 \qquad k = \omega/c \tag{105}$$

that satisfy the boundary condition at the fluid structure interface:

$$\partial p/\partial n = \rho\omega^2 w \tag{106}$$

where p is the fluid pressure, ρ is the density of the medium, and w is the normal component of the displacement of the surface of the structure. The pressure in the field can be written as

$$p = p_{\mathrm{I}} = p_{\mathrm{sc}} \tag{107}$$

where p is the total field pressure, p_{I} is the pressure in the incident wave, and p_{sc} is the pressure in the scattered wave. For points p on the structural surface S, the scattered pressure can be written in terms of the Helmholtz relation:

$$p_{\mathrm{sc}}(P) = \frac{1}{2\pi}\int_s \left[p(s)\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) - \frac{e^{ikr}}{r}\frac{\partial p(s)}{\partial n}\right] ds \tag{108}$$

The equation of the elastic structure can be written in operator form as:

$$L(w) = p(s) \tag{109}$$

where L is a differential operator. Equation (108) is an integral equation for the pressure. Equations (106–109) constitute the set of equations needed to solve for the scattered pressure. The integral equation (108) is solved by coupling Eqs. (106), (107), and (109) with Eq. (108), dividing the surface into many parts, and solving the resulting system of equations.

An alternative method of solution is offered by the asymptotic approximations which give the scattered pressure explicitly in terms of the motion of the surface. First, write the equation of motion of the elastic structure in matrix form as follows:

$$\begin{aligned} M_{\mathrm{s}}\ddot{x} + C_{\mathrm{s}}\ddot{x} + K_{\mathrm{s}}x &= f_{\mathrm{int}} + f_{\mathrm{ext}} \\ f_{\mathrm{ext}} &= -GA_{\mathrm{f}}(p_{\mathrm{I}} + p_{\mathrm{sc}}) \\ G^{\mathrm{T}}x &= u_{\mathrm{I}} + u_{\mathrm{sc}} \end{aligned} \tag{110}$$

where x is the structural displacement vector; M_{s}, C_{s}, and K_{s} are the structural mass, damping, and stiffness matrices, respectively; A_{f} is a diagonal area matrix for the fluid-structure interface; G is a transformation matrix that relates the forces on the structure to those on the interface; and f_{int} is the known internal force vector. The terms p_{I} and u_{I} are the (known) free-field pressure and fluid particle velocity associated with the incident wave, and p_{sc} and u_{sc} are the pressure and fluid particle velocity for the scattered wave. The dots denote differentiation with respect to time. The following fluid-structure interaction equations are then used to relate the pressure and motion on the fluid-structure interface.

1. First doubly asymptotic approximation (DAA1):

$$M_{\mathrm{f}}\dot{p}_{\mathrm{s}} + \rho c A_{\mathrm{f}} p_{\mathrm{s}} = \rho c \dot{M}_{\mathrm{f}}\left(G^{\mathrm{T}}\ddot{x} - \dot{u}_{\mathrm{I}}\right) \tag{111}$$

2. Second doubly asymptotic approximation (DAA2):

$$\begin{aligned} M_{\mathrm{f}}\ddot{p}_{\mathrm{s}} + \rho c A_{\mathrm{f}}\dot{p}_{\mathrm{s}} = \rho c\Omega_{\mathrm{f}}A_{\mathrm{f}}p_{\mathrm{s}} = \rho c\big[M_{\mathrm{f}}\left(G^{\mathrm{T}}\ddot{x} + \ddot{u}_I\right) \\ + \Omega_{\mathrm{f}}M_{\mathrm{f}}\left(G^{\mathrm{T}}\ddot{x} - \dot{u}_I\right)\big] \end{aligned} \tag{112}$$

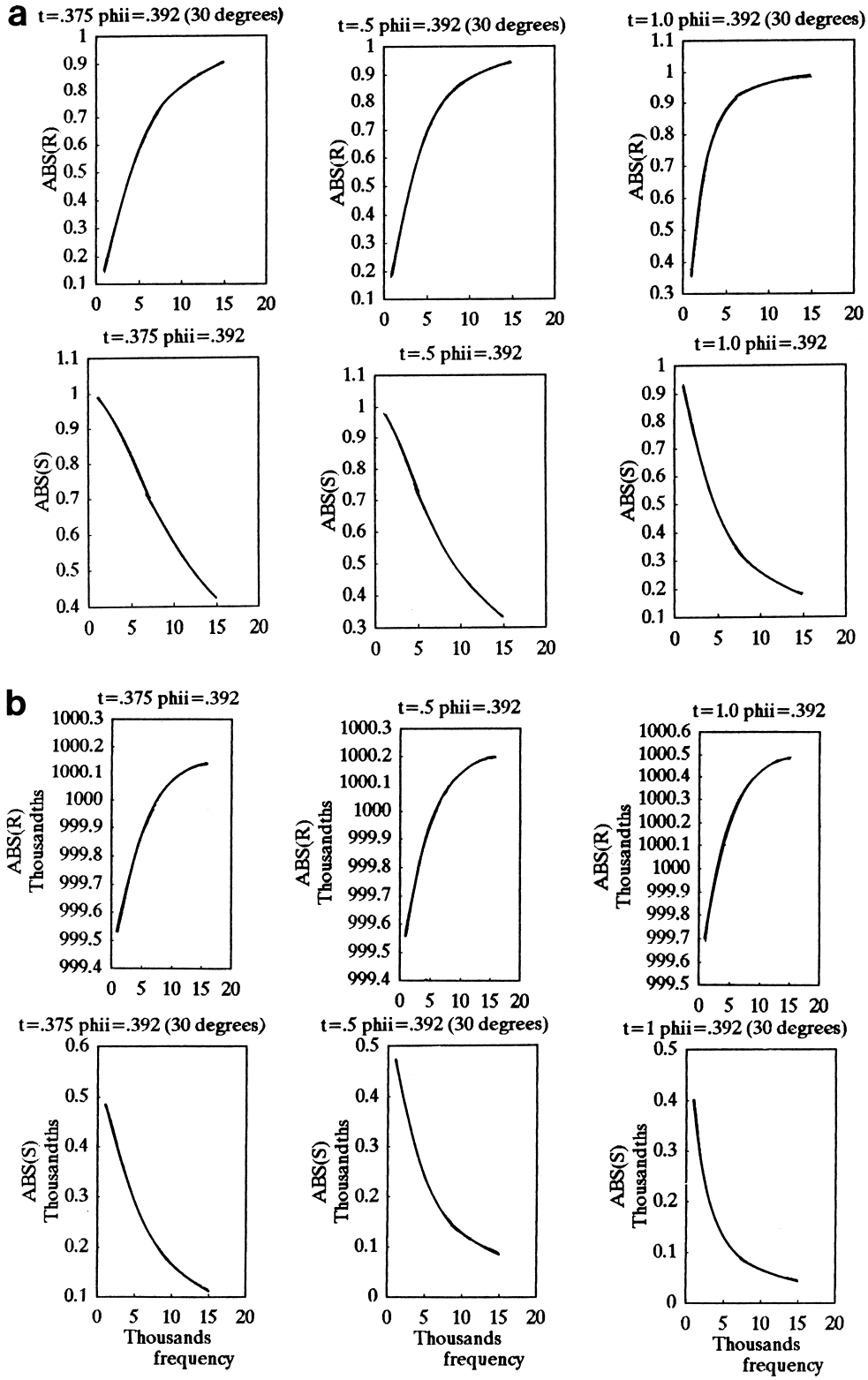


FIGURE 18 (a) Reflection (R) and transmission (S) coefficients for $\frac{3}{8}$-in., $\frac{1}{2}$-in., and 1-in. steel plates with water on both sides ($\phi_i = 30°$). (b) Reflection (R) and transmission (S) coefficients for $\frac{3}{8}$-in., $\frac{1}{2}$-in., and 1-in. steel plates with water on the incident wave side and air on the other side ($\phi_i = 30°$). (Note that ordinates are in thousandths; thus, $R \approx 1$ and S is very small.)

The M_f and Ω_f are the fluid added mass and frequency matrices pertaining to the fluid-structure interface, and G^T is the transpose of G.

In essence, the doubly asymptotic approximations uncouple the fluid–structure interaction by giving an explicit relation between pressure and motion of the surface. An illustration of the uncoupling procedure of the asymptotic approximations can be formulated by taking the very simplest case of a plane wave. The pressure and velocity are related by (this holds for high frequency):

$$p = \rho_o c u \tag{113}$$

where p is the pressure and u is the velocity. If this pressure were applied to a simple elastic plate, the resulting differential equation would be (noting that $u = \dot{w}$):

$$D\nabla^4 w + \rho h \ddot{w} = -\rho_o c \dot{w} \tag{114}$$

Thus, for this simple case, the entire fluid-structure interaction can be solved by one equation instead of having to solve the Helmholtz integral equation coupled with the elastic plate equation.

IX. RANDOM LINEAR ACOUSTICS

A. Linear System Equations

1. Impulse Response

Consider a system with n inputs and m outputs as shown in Fig. 19. Each of the inputs and outputs is a function of time. The central problem lies in trying to determine the outputs or some function of the outputs in terms of the inputs or some function of them. Let any input $x(t)$ be divided into a succession of impulses as shown in Fig. 20. Let $h(t-\tau)$ be the response at time t due to a unit impulse at time τ. A unit impulse is defined as one in which the area under the input versus time curve is unity. Thus, if

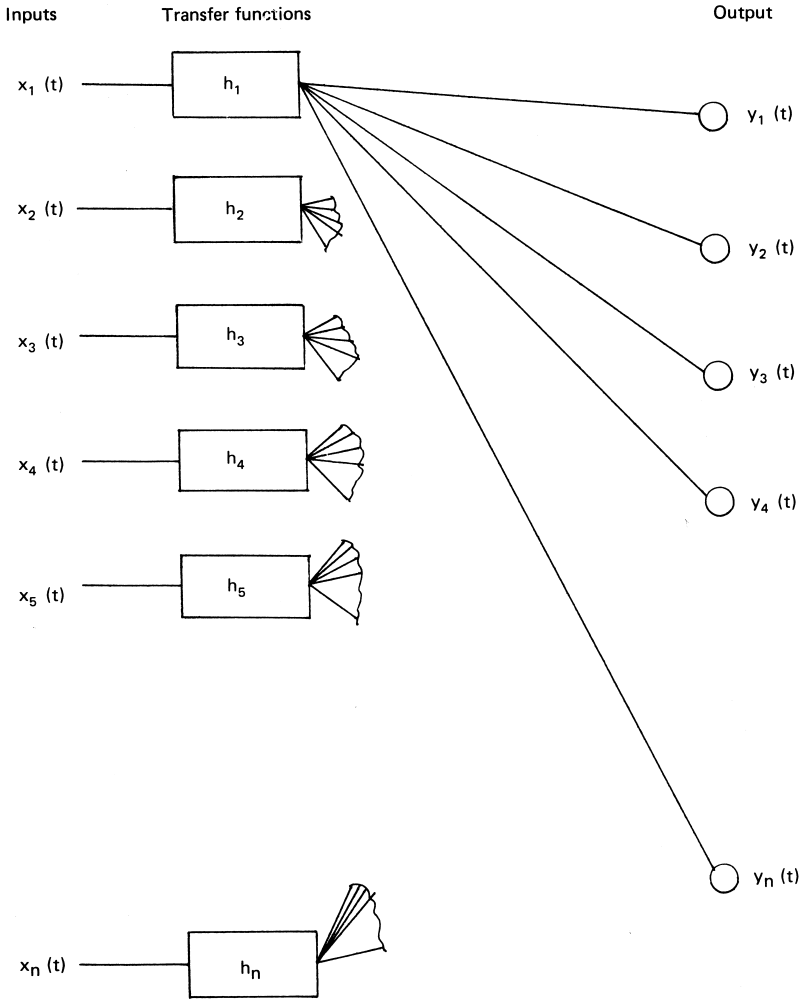


FIGURE 19 Block diagram for the system. $x_1(t)$, $x_2(t) \ldots x_n(t) = n$ inputs; h_1, h_2, $\ldots h_n = n$ transfer functions; $y_1(t)_1 t_2(t) \ldots y_m(t) = m$ outputs.

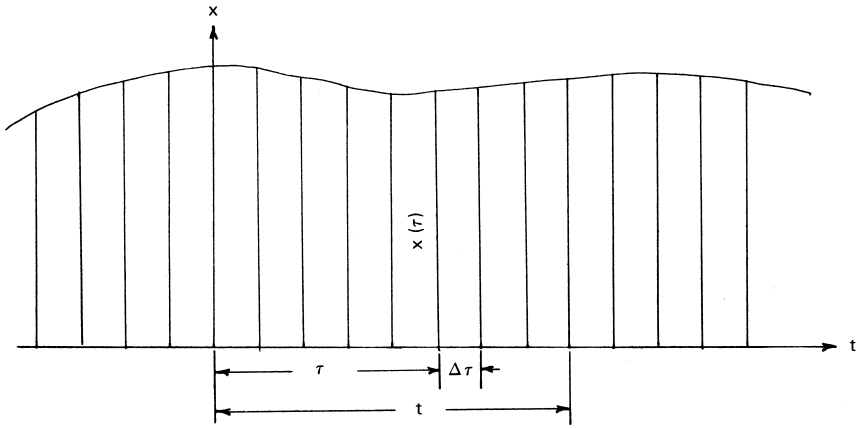


FIGURE 20 Input divided into infinitesimal impulses. t = time; τ = the value of time at which $x(\tau)$ is taken; $\Delta\tau$ = time width of impulse; $x(\tau)$ = value of the impulse at time τ.

the base is $\Delta\tau$, the height of the unit impulse is $1/\Delta\tau$. Thus, $h(t-\tau)$ is the response per unit area (or per unit impulse at $t=\tau$). The area (or impulse) is $x(\tau)\Delta\tau$. The response at time t is the sum of the responses due to all the unit impulses for all time up to t, that is, from $-\infty$ to t. But it is physically impossible for a system to respond to anything but past inputs; therefore,

$$h(t-\tau)=0 \qquad \text{for } \tau > t \tag{115}$$

Thus, the upper limit of integration can be changed to $+\infty$. By a simple change of variable ($\theta = t-\tau$) it can be demonstrated that:

$$y(t)=\int_{-\infty}^{+\infty} h(\theta)x(t-\theta)\,d\theta \tag{116}$$

Since there are n inputs and m outputs, there has to be one of these equations for each input and output. Thus, for the ith input and jth output,

$$y_{ij}(t)=\int_{-\infty}^{+\infty} h_{ij}(\tau)x_i(t-\tau)\,d\tau \tag{117}$$

2. Frequency Response Function

The frequency response function or the transfer function (the system function, as it is sometimes known) is defined as the ratio of the complex output amplitude to the complex input amplitude for a steady-state sinusoidal input. (The frequency response function is the output per unit sinusoidal input at frequency ω.) Thus, the input is

$$x_i(t)=\bar{x}_i(\omega)e^{i\omega t} \tag{118}$$

and the corresponding output is

$$y_j=\bar{y}_j(\omega)e^{i\omega t} \tag{119}$$

where $\bar{x}_i(\omega)$ and $\bar{y}_j(\omega)$ are the complex amplitudes of the input and output respectively. Then the frequency response function $H_{ij}(\omega)$ is

$$H_{ij}(\omega)=\frac{\bar{y}_j(\omega)}{\bar{x}_i(\omega)} \tag{120}$$

For sinusoidal input and output, Eq. (117) becomes:

$$\frac{\bar{y}_i(\omega)}{\bar{x}_i(\omega)}=\int_{-\infty}^{+\infty} h_{ij}(\tau)e^{-i\omega\tau}\,d\tau \tag{121}$$

It is therefore proven that the frequency response function is the Fourier transform of the unit impulse function.

3. Statistics of the Response

Since the linear process is assumed to be random, the results are based on statistical operations on the process. In this section, the pertinent statistical parameters will be derived. Referring back to Eq. (117), we see that the total response y_j is the sum over all inputs. Thus,

$$y_j(t)=\sum_{i=1}^{n}\int_{-\infty}^{+\infty} h_{ij}(\tau)x_i(t-\tau)\,d\tau \tag{122}$$

The cross correlation between outputs $y_j(t)$ and $y_k(t)$ is defined as follows:

$$C_{jk}(\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} y_j(t)y_k(t+\tau)\,dt \tag{123}$$

From the definition of $C_{jk}(\tau)$ it is seen that:

$$C_{kj}(\tau)=\langle y_k(t)y_j(t+\tau)\rangle$$

where:

$$\langle(\,)\rangle=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T}(\,)\,\mathrm{dt}$$

Substituting Eq. (122) and rearranging,

$$C_{jk}(\tau) = \sum_{s=1}^{m}\sum_{r=1}^{n}\int_{-\infty}^{+\infty} du \int_{-\infty}^{+\infty} dv \left\{ h_{js}(u)h_{kr}(v) \times \left[\lim_{Z\to\infty} \frac{1}{2T}\int_{-T}^{+T} x_s(t-u)x_r(t-v+\tau)\,dt \right] \right\} \tag{124}$$

By definition of the cross correlation,

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{+T} x_s(t-u)x_r(t-v+\tau)\,dt = C_{rs}(u-v+\tau) \tag{125}$$

Thus,

$$C_{jk}(\tau) = \sum_{s=1}^{m}\sum_{r=1}^{n}\int_{-\infty}^{+\infty} du \int_{-\infty}^{+\infty} dv \times [h_{js}(u)h_{kr}(v)C_{rs}(u-v+\tau)] \tag{126}$$

The cross spectrum $G_{jk}(\omega)$ is defined as the Fourier transform of the cross correlation. The inverse Fourier transform relation is, then,

$$C_{kj}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} G_{jk}(\omega)e^{i\omega\tau}\,d\omega$$

Thus,

$$G_{jk}(\omega) = \int_{-\infty}^{+\infty} C_{kj}(\tau)e^{-i\omega\tau}\,d\tau \tag{127}$$

Note that:

$$G_{kj}(\omega) = \int_{-\infty}^{+\infty} C_{kj}(\tau)e^{i\omega\tau}\,d\tau = \int_{-\infty}^{+\infty} C_{jk}(-\tau)e^{-i\omega\tau}\,d\tau = \int_{-\infty}^{+\infty} C_{jk}(\theta)e^{i\omega\theta}\,d\theta = G^*_{jk}(\omega)$$

where G^*_{jk} is the complex conjugate of G_{jk}.

Substituting Eq. (126), changing variables $\theta = u - v + \tau$, and using definition (127) and relation (121),

$$G_{jk}(\omega) = \sum_{s=1}^{n}\sum_{r=1}^{n} H^*_{js}(\omega)H_{kr}(\omega)G_{rs}(\omega) \tag{128}$$

in which H^*_{js} is the complex conjugate of H_{js}. Equation (128) gives the cross spectrum of the outputs $G_{jk}(\omega)$ in terms of the cross spectrum of the inputs $G_{rs}(\omega)$. In matrix notation, Eq. (128) can be written as:

$$G^{\mathrm{o}}(\omega) = H^* G^{\mathrm{i}} H^{\mathrm{T}} \tag{129}$$

where G^{o} is the output matrix of cross spectra, G^{i} is the input matrix of cross spectra, and H is the matrix of transfer functions. H^{T} denotes the transpose matrix of H, and H^* is the complex conjugate matrix of H. Thus,

$$G^{\mathrm{o}}(\omega) = \begin{vmatrix} G^{\mathrm{o}}_{11}(\omega) & G^{\mathrm{o}}_{12}(\omega) & G^{\mathrm{o}}_{13}(\omega) & \cdots & G^{\mathrm{o}}_{1k}(\omega) \\ G^{\mathrm{o}}_{21}(\omega) & & & & \\ \vdots & & & & \\ G^{\mathrm{o}}_{j1}(\omega) & \cdots & & & G^{\mathrm{o}}_{jj}(\omega) \end{vmatrix} \tag{130}$$

By virtue of the fact that $G_{ij}(\omega) = G^*_{ji}(\omega)$, the above matrix is a square Hermitian matrix. The input matrix G^{i} is

$$G^{\mathrm{i}}(\omega) = \begin{vmatrix} G^{\mathrm{i}}_{11}(\omega) & G^{\mathrm{i}}_{12}(\omega) & G^{\mathrm{i}}_{13}(\omega) & \cdots & G^{\mathrm{i}}_{15}(\omega) \\ G^{\mathrm{i}}_{21}(\omega) & & & & \\ \vdots & & & & \\ G^{\mathrm{i}}_{r1}(\omega) & \cdots & & & G^{\mathrm{i}}_{rr}(\omega) \end{vmatrix} \tag{131}$$

G_{i} is also Hermitian of order $r \times r = s \times s = r \times s$. The transfer function matrix is a $k \times r$ complex matrix (not Hermitian):

$$H(\omega) = \begin{vmatrix} H_{11}(\omega) & H_{12}(\omega) & \cdots & H_{1r}(\omega) \\ H_{21}(\omega) & & & \\ \vdots & & & \\ H_{k1}(\omega) & \cdots & & H_{kr}(\omega) \end{vmatrix} \tag{132}$$

4. Important Quantities Derivable from the Cross Spectrum

The cross spectrum can be used as a starting point to derive several important quantities. The spectrum of the response at point j is obtained by letting $k = j$ in Eq. (128). The autocorrelation is obtained by letting $k = j$ in Eq. (123). The mean-square response is further obtained from the autocorrelation by letting $\tau = 0$. If the Fourier inverse of Eq. (127) is used to determine mean square, then:

$$C_{jj}(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} G_{jj}(\omega)\,d\omega = \text{meam square} = M_j^2 = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} y_j^2(t)\,dt \tag{133}$$

If the mean square is desired over a frequency band $\Delta\Omega = \Omega_2 - \Omega_1$, then it is given by:

$$\left(M_j^2\right)_{\Delta\Omega} = \frac{1}{2\pi}\int_{\Omega_2} G_{jj}(\omega)\,d\omega \tag{134}$$

The mean value of $y_j(t)$ is defined as:

$$\bar{M}_j = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} y_j(t)\, dt \tag{135}$$

The variance σ_j^2 is defined as the mean square value about the mean:

$$\sigma_j^2 = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} [y_j(t) - \bar{M}_j]^2\, dt \tag{136}$$

The square root of the variance is known as the standard deviation. By using Eqs. (133) and (135), Eq. (136) can be written:

$$\sigma_j^2 = M_j^2 - (\bar{M}_j)^2 \tag{137}$$

The mean, the variance, and the standard deviation are three important parameters involved in probability distributions. Note that if the process is one with zero mean value, then the variance is equal to the mean square, and the standard deviation is the root mean square.

The above quantities are associated with the ordinary spectrum rather than the cross spectrum. An important physical quantity associated with the cross spectrum is the coherence, which is defined as:

$$\gamma_{jk}^2(\omega) = \frac{|G_{jk}(\omega)|^2}{G_{jj}(\omega) G_{kk}(\omega)} \tag{138}$$

The lower limit of γ_{jk}^2 must be zero since the lower limit of $G_{jk}(\omega)$ is zero. This corresponds to no correlation between the signals at j and k. In addition, $\gamma_{jk}^2 \le 1$. Going back to Eq. (128), we see that if there is only one input, then:

$$G_{jk}(\omega) = H_{js}^* H_{kr} G_{rr} \tag{139}$$

Thus,

$$\gamma_{jk}^2 = \frac{H_{js}^* H_{kr} H_{js} H_{kr}^* G_{rr}^2}{H_{js}^* H_{js} G_{rr} H_{kr}^* H_{kr} G_{rr}} = 1 \tag{140}$$

So the field is completely coherent for a single input to the system.

In an acoustic field, sound emanating from a single source is coherent. If the coherence is less than unity, then the field is partially coherent. The partial coherence effect is sometimes due to the fact that the source is of finite extent. It is also sometimes due to the fact that there are several sources causing the radiation and these sources are correlated in some way with each other.

5. The Cross Spectrum in Terms of Fourier Transforms

The cross spectrum can also be expressed in terms of Fourier transforms alone. To see this, start with the basic definition of cross spectrum as given by Eq. (127), where:

$$C_{jk}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} y_j(t) y_k(t+\tau)\, dt \tag{141}$$

Thus,

$$G_{jk}(\omega) = \int_{-\infty}^{+\infty} \left\{ \lim_{T\to\infty} \frac{1}{2T} \times \int_{-T}^{+T} y_j(t) y_k(t+\tau)\, dt \right\} e^{-i\omega\tau}\, d\tau \tag{142}$$

Letting $t = u$ and $t + \tau = \upsilon$, we have:

$$G_{jk} = \int_{-\infty}^{+\infty} \left\{ \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} y_j(u) y_k(\upsilon)\, du \right\} e^{-i\omega(\upsilon-u)}\, d\upsilon \tag{143}$$

$$= \int_{-\infty}^{+\infty} \left\{ \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} y_j(u) e^{i\omega u}\, du \right\} y_k(\upsilon) e^{-i\omega\upsilon}\, d\upsilon \tag{144}$$

The next step is true only under the condition that the process is ergodic. In this case, the last equation can be written as:

$$G_{jk}(\omega) = \lim_{T\to\infty} \frac{1}{2T} \left(\int_{-T}^{+T} y_j(u) e^{i\omega u}\, du \right) \times \left(\int_{-T}^{+T} y_k(\upsilon) e^{-i\omega\upsilon}\, d\upsilon \right) \tag{145}$$

This last relation can then be written:

$$G_{jk}(\omega) = \lim_{T\to\infty} \frac{\bar{y}_j^*(T,\omega) \bar{y}_k(T,\omega)}{2T} \tag{146}$$

where:

$$\bar{y}_j(T,\omega) = \int_{-T}^{+T} y_j(t) e^{i\omega t}\, dt \tag{147}$$

and y_j^* is the complex conjugate of y_j.

$$\bar{y}_k(T,\omega) = \int_{-T}^{+T} y_k(t) e^{-i\omega t}\, dt \tag{148}$$

Equation (146) expresses the cross spectrum in terms of the limit of the product of truncated Fourier transforms.

6. The Conceptual Meaning of Cross Correlation, Cross Spectrum, and Coherence

Given two functions of time $x(t)$ and $y(t)$, the cross correlation between these two functions is defined mathematically by the formula:

$$C(x, y, \tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} x(t) y(t+\tau)\, dt \tag{149}$$

This formula states that we take x at any time t, multiply it by y at a time $t+\tau$ (i.e., at a time τ later than t), and sum the product over all values ($-T < t < +T$). The result is then divided by $2T$. In real systems, naturally, T is finite, and the meaning of ∞ in the formula is that various values of T must be tried to make sure that the same answer results independent of T.

For two arbitrary functions of time, formula (149) has no real meaning. It is only when the two signals have something to do with each other that the cross correlation tells us something. To see this point clearly, consider an arbitrary random wave train moving in space. (It could be an acoustic wave, an elastic wave, an electromagnetic wave, etc.) Let $x(t) = x_1(t)$ be the response (pressure, stress, etc.) at one point, and $y(t) = x_2(t)$ be the response at another point. Now form the cross correlation between x_1 and x_2 (the limit is eliminated, it being understood):

$$C(x_1, x_2, \tau) = \frac{1}{2T}\int_{-T}^{+T} x_1(t)x_2(t+\tau)\,dt \qquad (150)$$

When the points coincide (i.e., $x_1 = x_2$), the relation becomes:

$$C(x_1, \tau) = \frac{1}{2T}\int_{-T}^{+T} x_1(t)x_1(t+\tau)\,dt \qquad (151)$$

and if $\tau = 0$,

$$C(x_1, 0) = \frac{1}{2T}\int_{-T}^{+T} x_1^2(t)\,dt \qquad (152)$$

which is, by definition, the mean square value of the response at point x_1. For other values of τ, Eq. (151) defines the autocorrelation at point 1. It is the mean value between the response at one time and the response at another time τ later than t. Thus, Eq. (150) is the mean product between the response at point 1 and the response at point 2 at a time τ later.

Now, going back to the random wave train, let us assume that it is traveling in a nondispersive medium (i.e., with velocity independent of frequency). It is seen that if the wave train leaves point 1 at time t (see Fig. 21) and travels through the system with no distortion, then:

$$y(t) = x_2(t) = Ax_1(t-\tau_1) \qquad (153)$$

where A is some decay constant giving the amount that the wave has decreased in amplitude from point 1 to point 2, and τ is the time of travel from 1 to 2. Forming the cross correlation $C(x_1, x_2, \tau)$ gives

$$C(x_1, x_2, \tau) = \frac{1}{2T}\int_{-T}^{+T} x_1(t)Ax_1(t+\tau-\tau_1)\,dt$$
$$= AC(x_1, x_1, \tau-\tau_1) \qquad (154)$$

Thus, the cross correlation of a random wave train in a nondispersive system is exactly the same form as the autocorrelation of the wave at the starting point, except that the peak occurs at a time delay corresponding to the time necessary for the wave to travel between the points. In the absence of attenuation, the wave is transmitted undisturbed in the medium. However, in most cases it is probable that the peak is attenuated somewhat as the wave progresses. It is thus seen that cross correlation is an extremely useful concept for measuring the time delay of a propagating random signal. In the above case, it had to be assumed that the signal was propagating in a nondispersive medium and that when the cross correlation was done, the signal was actually being traced as it moved through the system.

Consider the meaning of cross correlation if the system was dispersive (i.e., if the velocity was a function of frequency). White has addressed himself to this question and has demonstrated that time delays in the cross correlation can still be measured with confidence if the signal that is traveling is band-limited noise. For dispersive systems where the velocity is a function of frequency, it has been pointed out in the literature that time delays can also be obtained. For this case, the following cross spectrum is formed:

$$S_{12}(\omega) = \int_{-\infty}^{+\infty} C_{12}(\tau)e^{-i\omega\tau}\,d\tau \qquad (155)$$

The cross spectrum is a complex number and can be written in terms of amplitude and phase angle $\theta_{12}(\omega)$ as follows:

$$S_{12}(\omega) = |S_{12}(\omega)|e^{-i\theta_{12}(\omega)} \qquad (156)$$

The phase angle $\theta_{12}(\omega)$ is actually the phase between input and output at frequency ω. The time delay from input to output is then:

$$\tau(\omega) = \theta_{12}(\omega)/\omega \qquad (157)$$

Suppose that the signal has lost its propagating properties in that it has reflected many times and set up

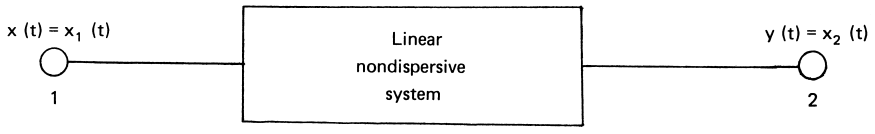


FIGURE 21 Input and output in a linear system. $x(t)$ = input; $y(t)$ + output.

a reverberant field in the system. Consider the physical meaning of cross correlation in this case. To answer this question partially, examine an optical field. In optical systems, extensive use has been made of the concept of partial coherence.

At the beginning of this section, two functions $x(t)$ and $y(t)$ were chosen, and the cross correlation between them was formed. It was pointed out that if the two functions are completely arbitrary, then there is no real physical meaning to the cross correlation. However, if the two functions are descriptions of response at points of a field, then there is a common ground to interpret cross correlation. Thus, the cross correlation and any function that is derived therefrom give some measure of the dependence of the vibrations at one point on the vibrations at the other point. This is a general statement, and to tie it down, the concept *coherence* has been used.

In the optical case, suppose light is coming into two points in a field. If the light comes from a small, single source of narrow spectral range (i.e., almost single frequency), then the light at the two field points is dependent. If each point in the field receives light from a different source, then the light at the two field points is independent. The first case is termed coherent, and the field at the points is highly correlated (or dependent). The second case is termed incoherent, and the field between the two points is uncorrelated (independent).

These are the two extreme cases, and between them there are degrees of coherence (i.e., partial coherence). Just as in everyday usage, coherence is analogous to clarity or sharpness of the field, whereas incoherence is tantamount to haziness or "jumbledness." The same idea is used when speaking about someone's speech or written article. If it is concise and presented clearly, it is coherent. If the ideas and presentation are jumbled, they can be called incoherent.

Single-frequency radiation is coherent radiation; radiation with a finite bandwidth is not. The partial coherence associated with finite spectral bandwidth is called the temporal (or timewise) coherence. On the other hand, light or sound emanating from a single source gives coherent radiation, but a point source is never actually obtained. The partial coherence effect, due to the fact that the source is of finite extent, is termed space coherence. The point source gives unit coherence in a system, whereas an extended source gives coherence somewhat less than unity.

The square of coherence $\gamma_{12}(\omega)$ between signals at points 1 and 2 at frequency ω is defined as:

$$\gamma_{12}^2(\omega) = \frac{|S_{12}(\omega)|^2}{S_{11}(\omega)S_{22}(\omega)} \tag{158}$$

where:

$$S_{12}(\omega) = \int_{-\infty}^{+\infty} C_{12}(\tau)e^{-i\omega\tau}\,d\tau \tag{159}$$

in which $C_{12}(\tau)$ is the cross correlation between signals at points 1 and 2. The function $S_{12}(\omega)$ is the cross spectrum between signals at points 1 and 2, and $S_{11}(\omega)$ and $S_{22}(\omega)$ are the autospectra of the signals at points 1 and 2, respectively. Wolf has other ways of defining coherence by functions called *complex degree of coherence or mutual coherence function*, but it all amounts conceptually to the same cross spectrum as given by Eq. (158).

Although there are formal proofs that $\gamma_{12}(\omega)$ is always between 0 and 1, one can reason this out nonrigorously by going back to the basic physical ideas associated with correlation and coherence. If the signals at two points are uncorrelated, and therefore incoherent, the cross correlation is zero, thus $\gamma_{12}(\omega)$ is zero. If the signals are perfectly correlated, then this is tantamount to saying that the signals in the field are a result of input to the system from a single source, as shown in Fig. 22.

As seen before, the cross spectrum $S_{yz}(\omega)$ can be written in terms of the input spectrum $S_x(\omega)$ and the transfer functions $Y_z(i\omega)$ and $Y_y(i\omega)$ as follows:

$$S_{yz}(i\omega) = Y_y(i\omega)Y_z^*(i\omega)S_x(\omega) \tag{160}$$

Thus,

$$\gamma_{yz}^2(\omega) = \frac{|S_{yz}(\omega)|^2}{S_{yy}(\omega)S_{zz}(\omega)} \tag{161}$$

$$= \frac{Y_y(i\omega)Y_z^*(i\omega)Y_y^*(i\omega)Y_z(i\omega)S_x^2(\omega)}{|Y_y(i\omega)|^2S_x(\omega)|Y_z(i\omega)|^2S_x(\omega)} = 1 \tag{162}$$

So that:

$$0 \le \gamma_{12}(\omega) \le 1 \qquad \text{(complete coherence)} \tag{163}$$

Between the cases of complete coherence and complete incoherence there are many degrees of *partial coherence*.

B. Statistical Acoustics

1. Physical Concept of Transfer Function

In Section IX.A.2 it was shown that H_{ij} was the transfer function that gave the output at j per unit sinusoidal input at i. Suppose there is an acoustic field which is generated by a group of sound sources and these sources are surrounded by an imaginery surface S_o as shown in Fig. 23.

Through each element of S_o, sound passes into the field. Thus, each element of S_o, denoted by ds, can be considered a source that radiates sound into the field. Consider the

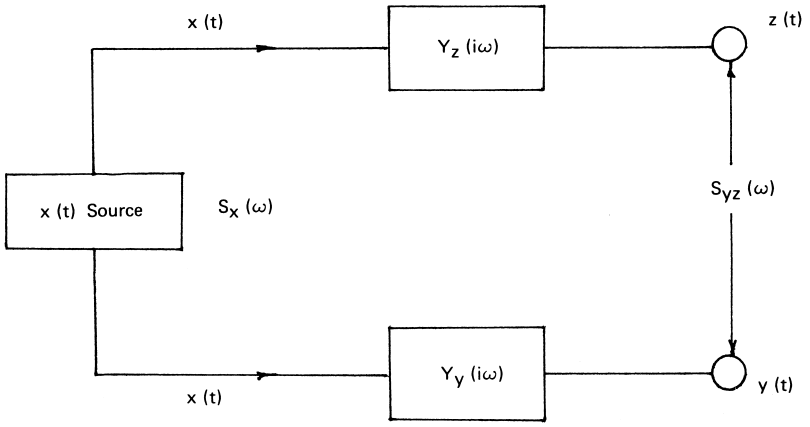


FIGURE 22 Single input or coherent system. $x(t)$ = input; $S_x(\omega)$ = spectrum of input; $Y_z(i\omega)$ = transfer function for z output; $Y_y(i\omega)$ = transfer function for y output; $z(t)$, $y(t)$ = outputs.

pressure $dp(P, \omega)$ at field point P at frequency ω due to radiation out of element ds,

$$dp(P, \omega) = H_p(P, S, \omega)p(S, \omega)\,ds \tag{164}$$

where $H_p(P, S, \omega)$ is the pressure at field point P per unit area of S due to a unit sinusoidal input pressure on S. The total pressure in the field at point P is

$$p(P, \omega) = \int_{S_o} H_p(P, S, \omega)p(S, \omega)\,ds \tag{165}$$

If motion (e.g., acceleration) of the surface S is considered instead of pressure, the counterpart to Eq. (165) is

$$p(P, \omega) = \int_{S_o} H_a(P, S, \omega)a(S, \omega)\,ds \tag{166}$$

where $H_a(P, S, \omega)$ is the transfer function associated with acceleration; that is, it is the pressure at field point P per unit area of S due to a unit sinusoidal input acceleration of S.

Applying these ideas to Eq. (128), it is seen that the cross spectrum of the field pressure can immediately be written

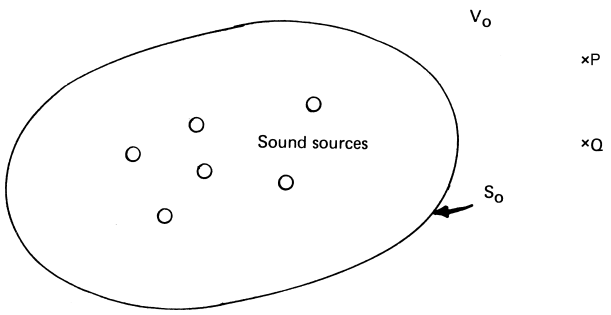


FIGURE 23 Surface surrounding the sources. S_o = surrounding surface; V_o = volume outside the surface; P, Q = field points.

in terms of the cross spectrum of the surface pressure or the cross spectrum of the surface acceleration. For surface pressure, Eq. (128) becomes:

$$G(P, Q, \omega) = \int_{S_i}\int_{S_r} H_p^*(P, S_i, \omega)H_p(Q, S_r, \omega) \times G(S_i, S_r, \omega)\,ds_i\,ds_r \tag{167}$$

Comparing Eq. (167) with Eq. (128) we find that the points j, k become field points P, Q. $G(P, Q, \omega)$ is the cross spectrum of pressure at field points P, Q. The transfer functions $H_{js}^*(\omega)$ become $H_p^*(P, S_i, \omega)$ in which S_i is the surface point or ith input point. $H_{kr}(\omega)$ becomes $H_p(Q, S_r, \omega)$ where S_r is the other surface point or rth input point. $G(S_i, S_r, \omega)$ is the input cross spectrum, which is the cross spectrum of surface pressure. The summations over r and s become integrals over S_i and S_r. For acceleration,

$$G(P, Q, \omega) = \int_{S_i}\int_{S_r} H_a^*(P, S_i, \omega)H_a(Q, S_r, \omega) \times A(S_i, S_r, \omega)\,ds_i\,ds_r \tag{168}$$

The transfer functions are those for acceleration, and $A(S_i, S_r, \omega)$ is the cross spectrum of the surface acceleration. The relation can be written for any other surface input such as velocity.

2. Response in Terms of Green's Functions

The Green's functions for single-frequency systems were taken up in a previous section. The transfer function for pressure is associated with the Green's function that vanishes over the surface. Thus,

$$H_p(P, S, \omega) = \frac{\partial g_1(P, S, \omega)}{\partial n} \tag{169}$$

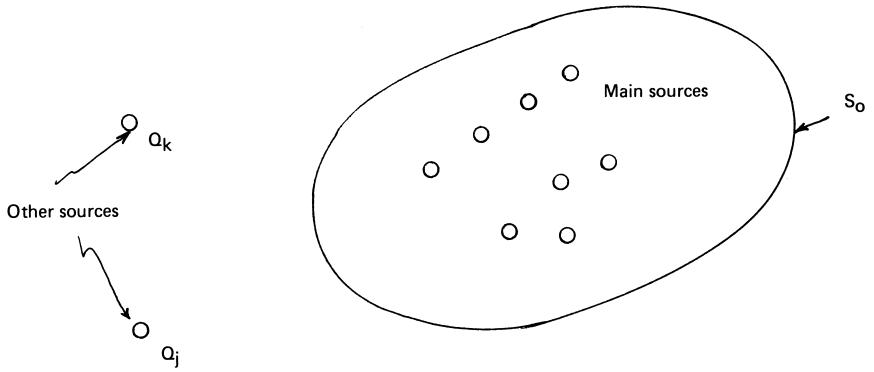


FIGURE 24 Surface surrounding main sources with presence of other field sources. S_o = surrounding surface; Q_k, Q_j = strengths of other sources not within S_o.

The transfer function for acceleration is associated with the Green's function whose normal derivative vanishes over S_o. Thus,

$$H_a(P, S, \omega) = \rho g_2(P, S, \omega) \tag{169a}$$

The statistical relations for the field pressure can therefore immediately be written in terms of the cross spectrum of pressure or acceleration over the surface surrounding the sources,

$$G(P, Q, \omega) = \int_{S_i}\int_{S_r} \frac{\partial g_1^*(P, S_i, \omega)}{\partial n_i}\frac{\partial g_1(Q, S_r, \omega)}{\partial n_r} \times G(S_i, S_r, \omega)\, ds_i\, ds_r \tag{170}$$

or

$$G(P, Q, \omega) = \int_{S_i}\int_{S_r} \rho^2 g_2^*(P, S_i, \omega) g_2(Q, S_r, \omega) \times A(S_i, S_r, \omega)\, ds_i\, ds_r \tag{171}$$

These relations give the cross spectrum of the field pressure as a function of either the cross spectrum of the surface pressure $G(S_i, S_r, \omega)$ or the cross spectrum of the surface acceleration $A(S_i, S_r, \omega)$. Equation (170) was derived by Parrent using a different approach. At this point, one should review the relationship between Eqs. (170) and (171) for acoustic systems and (128) for general linear systems. It is evident that the inputs in Eqs. (170) and (171) are $G(S_i, S_r, \omega)$, $A(S_i, S_r, \omega)$, respectively, and the output is $G(P, Q, \omega)$ in both cases. The frequency response functions are the transfer functions described by Eqs. (169) and (169a).

3. Statistical Differential Equations Governing the Sound Field

Consider the general case where there are source terms present in the field equation. This is tantamount to saying that outside the series of main radiating sources which have been surrounded by a surface (see Fig. 24) there are other sources arbitrarily located in the field. For example, in the case of turbulence surrounding a moving structure, the turbulent volume constitutes such a source, whereas the surface of the structure surrounds all the other vibrating sources. The equation governing the propagation of sound waves in the medium is

$$\nabla^2 p(P, t) - \frac{1}{c_0^2}\frac{\partial^2 p(P, t)}{\partial t^2} = V(Q, t) \tag{172}$$

where:

$$V(Q, t) = \sum_i V_i(Q_i, t) \tag{173}$$

In the above equation, V is a general source term that may consist of a series of sources at various points in the medium. Actually, the medium being considered is bounded internally by S_o so that the sources inside S_o are not in the medium. The sources at Q_i, however, are in the medium.

The various types of source terms that can enter acoustical fields arise from the injection of mass, momentum, heat energy, or vorticity into the field. These are discussed by Morse and Ingard and will not be treated here. It is assumed that the source term $V(Q, t)$ is a known function of space and time, or if it is a random function, then some statistical information is known about it such as its cross correlation or cross spectrum.

In cases where the field is random, a statistical description has to be used. The cross correlation function $\Gamma(P_1, P_2, \tau)$ between pressures at field points P_1 and P_2 are defined as:

$$\Gamma(P_1, P_2, \tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} p(P_1, t)p(P_2, t+\tau)\, dt \tag{174}$$

(To the author's knowledge, one of the first pieces of work on correlation in wavefields was the paper of Marsh.)

The Fourier transform, $U(P, \omega)$ of the pressure $p(P, t)$ is

$$U(P, \omega) = \int_{-\infty}^{+\infty} p(P, t) e^{-i\omega t}\, dt \tag{175}$$

Taking the inverse, we can write the above equation

$$\nabla^2 p = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \nabla^2 U(P, \omega) e^{i\omega t}\, d\omega \tag{176}$$

Also from the Fourier transform of $V(Q, t)$:

$$W(Q, \omega) = \int_{-\infty}^{+\infty} V(Q, t) e^{-i\omega t}\, dt \tag{177}$$

and its inverse:

$$V(Q, t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} W(Q, \omega) e^{i\omega t}\, d\omega \tag{178}$$

Substitution into the original nonhomogeneous wave equation (172) gives:

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \left[\nabla^2 U + \frac{\omega^2}{c_0^2} U - W \right] d\omega = 0 \tag{179}$$

Thus, for this relation to hold for all P and all ω, there must be

$$\nabla^2 U(P, \omega) + k^2 U(P, \omega) = W(Q, \omega) \tag{180}$$

The cross spectrum $G(P_1, P_2, \omega)$ between the pressures at P_1 and P_2 at frequency ω is defined in terms of the cross correlation $\Gamma(P_1, P_2, \tau)$, by:

$$G(P_1, P_2, \omega) = \int_{-\infty}^{+\infty} \Gamma(P_1, P_2, \tau) e^{i\omega\tau}\, d\tau \tag{181}$$

and the inverse is

$$\Gamma(P_1, P_2, \tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} G(P_1, P_2, \omega) e^{-i\omega\tau}\, d\omega \tag{182}$$

Thus, $\Gamma(P_1, P_2, \tau)$ can be written as:

$$\Gamma(P_1, P_2, \tau) = \lim_{T\to\infty} \frac{1}{2T} \times \int_{-T}^{+T} \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} U(P_1, \omega) e^{-i\omega t}\, d\omega \right] \times \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} U(P_2, \omega) e^{-i\omega(t+\tau)}\, d\omega \right] dt \tag{183}$$

So

$$\nabla_2^2 \Gamma(P_1, P_2, \tau) = \lim_{T\to\infty} \frac{1}{2T} \times \int_{-T}^{+T} \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} U(P_1, \omega) e^{-i\omega t}\, d\omega \right] \times \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} \nabla_2^2 U(P_2, \omega) e^{-i\omega(t+\tau)}\, d\omega \right] dt \tag{184}$$

where ∇_2^2 stands for operations performed in the $P_2(x_2, y_2, z_2)$ coordinates. Also,

$$\frac{\partial^2 \Gamma(P_1, P_2, \tau)}{\partial \tau^2} = \lim_{T\to\infty} \frac{1}{2T} \times \int_{-T}^{+T} \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} U(P_1, \omega) e^{-i\omega t}\, d\omega \right] \times \left[\frac{1}{2\pi} \int_{-\infty}^{+\infty} (-\omega^2) U(P_2, \omega) e^{-i\omega(t+\tau)}\, d\omega \right] dt \tag{185}$$

Thus,

$$\nabla_2^2 \Gamma(P_1, P_2, \tau) - \frac{1}{c_0^2} \frac{\partial^2 \Gamma(P_1, P_2, \tau)}{\partial \tau^2} = \langle p(P_1, t) V(Q_2, t + \tau) \rangle \tag{186}$$

It should be clear from an analysis similar to that given above that the following relation also holds:

$$\nabla_1^2 \Gamma(P_2, P_1, \tau) - \frac{1}{c_0^2} \frac{\partial^2 \Gamma(P_2, P_1, \tau)}{\partial \tau^2} = \langle p(P_2, t) V(Q_1, t + \tau) \rangle \tag{187}$$

This set of Eqs. (186) and (187) is an extension of the equation obtained by Eckart and Wolf. If the source term were zero, then:

$$\begin{aligned} \nabla_2^2 \Gamma(P_1, P_2, \tau) - \frac{1}{c_0^2} \frac{\partial^2 \Gamma(P_1, P_2, \tau)}{\partial \tau^2} &= 0 \\ \nabla_1^2 \Gamma(P_2, P_1, \tau) - \frac{1}{c_0^2} \frac{\partial^2 \Gamma(P_2, P_1, \tau)}{\partial \tau^2} &= 0 \end{aligned} \tag{188}$$

However, since:

$$\Gamma(P_2, P_1, \tau) = \Gamma(P_1, P_2, -\tau) \tag{189}$$

and

$$\frac{\partial^2}{\partial \tau^2} = \frac{\partial^2}{\partial(-\tau)^2} \tag{190}$$

Then,

$$\nabla_{1,2}^2 \Gamma(P_1, P_2, \tau) - \frac{1}{c_0^2} \frac{\partial^2 \Gamma(P_1, P_2, \tau)}{\partial \tau^2} = 0 \tag{191}$$

From the above relations, it is seen that the cross correlation is propagated in the same way that the original pressure wave propagates, except that real time t is replaced by correlation time τ.

The nonhomogeneous counterparts given by Eqs. (186) and (187) state that the source term takes the statistical

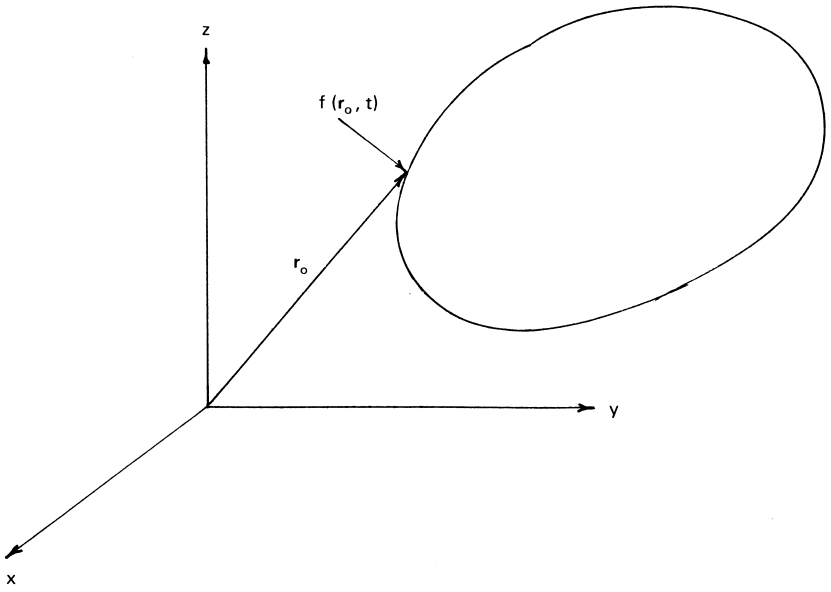


FIGURE 25 The loaded structure. $f(r_o, t)$ = force component at location $\mathbf{r}_o$ and time t; $w(\mathbf{r}, t)$ = deflection component at $\mathbf{r}$ at time t.

form of the cross correlation between the pressure p at a reference point and the source function V. Taking the Fourier transform of Eqs. (186) and (187), we see that the cross spectrum satisfies:

$$\nabla_2^2 G(P_1, P_2, \omega) + k^2 G(P_1, P_2, \omega) = \Phi_2(P_1, Q_2, \omega)$$

$$\nabla_1^2 G(P_2, P_1, \omega) + k^2 G(P_2, P_1, \omega) = \Phi_1(P_2, Q_1, \omega) \tag{192}$$

where Φ_1 and Φ_2 are the Fourier transforms of the cross correlation between the reference pressure and source function, that is

$$\Phi_2(P_1, Q_2, \omega) = \int_{-\infty}^{+\infty} \langle p(P_1, t) V(Q_2, t+\tau) \rangle e^{-i\omega\tau}\, d\tau$$

$$\Phi_1(P_2, Q_1, \omega) = \int_{-\infty}^{+\infty} \langle p(P_2, t) V(Q_1, t+\tau) \rangle e^{-i\omega\tau}\, d\tau \tag{193}$$

Thus, Φ_1 and Φ_2 are cross-spectrum functions between the pressure and source term.

In Eqs. (186), (187), (192), and (193), it is important to note that one point is being used as a reference point and the other is the actual variable. For example, in Eq. (187), the varying is being done in the $P_1(x_1, y_1, z_1)$ coordinates; thus, all cross correlations are performed with P_2 fixed. Conversely, in Eq. (186) all the operations are being carried out in the P_2 space with P_1 remaining fixed.

C. Statistics of Structures

1. Integral Relation for the Response

Let the loading (per unit area) on the structure be represented by the function $f(\mathbf{r}_0, t)$ where $\mathbf{r}_0$ is the position vector of a loaded point on the body with respect to a fixed system of axes, as shown in Fig. 25. Let the unit impulse response be $h(\mathbf{r}, \mathbf{r}_0, t-\theta)$; this is the output at $\mathbf{r}$ corresponding to a unit impulse at $t = 0$ and at location $\mathbf{r}_0$. The response at $\mathbf{r}$ at time t due to an arbitrary distributed excitation $f(\mathbf{r}_0, t)$ can then be written:

$$w(\mathbf{r}, t) = \int_{\mathbf{r}_0} d\mathbf{r}_0 \int_{-\infty}^{t} f(\mathbf{r}_0, \theta) h(\mathbf{r}, \mathbf{r}_0, t-\theta)\, d\theta \tag{194}$$

The integration is taken over the whole loaded surface denoted by $\mathbf{r}_0$. Since the loading is usually random in nature, only the statistics of the response (that is, the mean square values, the power spectral density, and so on) are determinable. Thus, let $U = t - \theta$ and form the cross correlation of the response at two points $\mathbf{r}_1$ and $\mathbf{r}_2$. This cross correlation is denoted by R_w and is

$$R_w(\mathbf{r}_1, \mathbf{r}_2, \tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} \left[\int_{\mathbf{r}_0} d\mathbf{r}_0 \int_{-\infty}^{+\infty} \times f(\mathbf{r}_0, t - U_1) h(\mathbf{r}_1, \mathbf{r}_0, U_1)\, dU_1 \right] \times \left[\int_{\mathbf{r}_0} d\mathbf{r}_0' \int_{-\infty}^{+\infty} f(\mathbf{r}_0', t - U_2 + \tau) \times h(\mathbf{r}_2, \mathbf{r}_0', U_2)\, dU_2 \right] dt \tag{195}$$

or

$$R_w(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \times h(\mathbf{r}_1, \mathbf{r}_0, U_1)h(\mathbf{r}_2, \mathbf{r}'_0, U_2)\, dU_1\, dU_2 \times \left[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{+T} f(\mathbf{r}_0, t-U_1) \times f(\mathbf{r}'_0, t-U_2+\tau)\, dt \right] \quad (196)$$

We assume a stationary process so that the loading is only a function of the difference of the times $t-U_2+\tau$ and $t-U_1$. Let

$$\begin{aligned}\tau_3 &= (t-U_2+\tau)-(t-U_1)\\ &= U_1-U_2+\tau\end{aligned} \quad (197)$$

then,

$$R_w(\mathbf{r}_1, \mathbf{r}_2, \tau) = \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \times h(\mathbf{r}_1, \mathbf{r}_0, U_1)(\mathbf{r}_2, \mathbf{r}'_0, U_2) \times R_\ell(\mathbf{r}_0, \mathbf{r}'_0, \tau_3)\, dU_1\, dU_2 \quad (198)$$

where $R_\ell(\mathbf{r}_0, \mathbf{r}'_0, \tau_3)$ is the cross-correlation function of the loading. Now form the cross spectrum of the response:

$$\begin{aligned}S_w(\mathbf{r}_1, \mathbf{r}_2, \omega) &= \int_{-\infty}^{+\infty} R_w(\mathbf{r}_1, \mathbf{r}'_2, \tau_3)e^{-i\omega\tau}\, d\tau\\ e^{-i\omega\tau} &= e^{i\omega(\tau_3-U_1+U_2)}\end{aligned} \quad (199)$$

Thus,

$$S_w(\mathbf{r}_1, \mathbf{r}_2, \omega) = \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 \int_{-\infty}^{+\infty} h(\mathbf{r}_1, \mathbf{r}_0, U_1)e^{i\omega U_1}\, dU_1 \times \int_{-\infty}^{+\infty} h(\mathbf{r}_2, \mathbf{r}'_0, U_2)e^{-i\omega U_2}\, dU_2 \times \left[\int_{-\infty}^{+\infty} R_\ell(\mathbf{r}_0, \mathbf{r}'_0, \tau_3)e^{-i\omega\tau_3}\, d\tau_3\right] \quad (200)$$

but the Fourier transform of the impulse function is the Green's function. Thus,

$$\begin{aligned}\int_{-\infty}^{+\infty} h(\mathbf{r}_1, \mathbf{r}_0, U_1)e^{i\omega U_1}\, dU_1 &= G^*(\mathbf{r}_1, \mathbf{r}_0, \omega)\\ \int_{-\infty}^{+\infty} h(\mathbf{r}_2, \mathbf{r}'_0, U_2)e^{-i\omega U_2}\, dU_2 &= G(\mathbf{r}_2, \mathbf{r}'_0, \omega)\end{aligned} \quad (201)$$

where G^* denotes the complex conjugate of G. The Green's function $G(\mathbf{r}_2, \mathbf{r}'_0, \omega)$ is the response at $\mathbf{r}_2$ due to a unit sinusoidal load at $\mathbf{r}'_0$, and $G^*(\mathbf{r}_1, \mathbf{r}_0, \omega)$ is the complex conjugate of the response at $\mathbf{r}_1$ due to a unit sinusoidal load at $\mathbf{r}_0$. The bracket can be written:

$$S_\ell(\mathbf{r}_0, \mathbf{r}'_0, \omega) = \int_{-\infty}^{+\infty} R_\ell(\mathbf{r}_0, \mathbf{r}'_0, \tau_3)e^{-i\omega\tau_3}\, d\tau_3 \quad (202)$$

where S_ℓ is the cross spectrum of the load. Thus, the expression for the cross spectrum of the response becomes:

$$S_w(\mathbf{r}_1, \mathbf{r}_2, \omega) = \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 G^*(\mathbf{r}_1, \mathbf{r}_0, \omega) \times G(\mathbf{r}_2, \mathbf{r}'_0, \omega)S_\ell(\mathbf{r}_0, \mathbf{r}'_0, \omega) \quad (203)$$

The spectrum at any point $\mathbf{r}_1$ is obtained by setting $\mathbf{r}_1 = \mathbf{r}_2$; thus,

$$S_w(\mathbf{r}_1, \omega) = \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 G^*(\mathbf{r}_1, \mathbf{r}_0, \omega) \times G(\mathbf{r}_1, \mathbf{r}'_0, \omega)S_\ell(\mathbf{r}_0, \mathbf{r}'_0, \omega) \quad (204)$$

Note the equivalence between Eq. (203) and the general Eq. (128) for linear systems.

In many practical cases, especially in turbulence excitation, the cross spectrum of the loading takes a homogeneous form as follows:

$$S_\ell(\mathbf{r}_0, \mathbf{r}'_0, \omega) = S(\mathbf{r}_0 - \mathbf{r}'_0, \omega) \quad (205)$$

Let $\mathbf{r}_0 - \mathbf{r}'_0 = \xi$. Equation (203) now becomes:

$$S_w(\mathbf{r}_1, \mathbf{r}_2, \omega) = \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 G(\mathbf{r}_1, \mathbf{r}'_0, \omega) \times G(\mathbf{r}_2, \mathbf{r}'_0, \omega)S(\xi, \omega) \quad (206)$$

White has shown that by applying Parseval's theorem and letting $\mathbf{r}_1 = \mathbf{r}_2$, the above equation can be written:

$$S_w(\mathbf{r}_1, \mathbf{r}'_1, \omega) = (2\pi)^2 \int_{\mathbf{k}} \tilde{S}(\mathbf{k}, \omega)\psi(\mathbf{k}, \omega)\, d\mathbf{k} \quad (207)$$

where:

$$\psi(\mathbf{k}, \omega) = \left|\frac{1}{(2\pi)^2}\int_{-\infty}^{+\infty} G(\mathbf{r}_1, \mathbf{r}, \omega)e^{ik(\mathbf{r}-\mathbf{r}_1)}\, d\mathbf{r}\right|^2 \quad (208)$$

In the above equations, $\tilde{S}(\mathbf{k}, \omega)$ is the spectrum of the excitation field in wave number space, and $\psi(\mathbf{k}, \omega)$ is the square of the Fourier transform of the Green's function, which can be obtained very quickly on a computer by application of the fast Fourier transform technique.

The Green's functions take on the true spatial character of an *influence function*. They represent the response at one point due to a unit sinuosidal load at another point. The inputs are loads, the outputs are deflections, and the linear *black boxes* are pieces of the structure as used in the first section of this chapter.

A few very interesting results can immediately be written from Eq. (204). Supposing a body is loaded by a single random force at point p, the loading $\ell(\mathbf{r}, t)$ can be written:

$$\ell(\mathbf{r}, t) = P(t)\delta(\mathbf{r} - \mathbf{r}_p) \quad (209)$$

The δ function signifies that ℓ is 0 except when $\mathbf{r} = \mathbf{r}_p$. Thus,

$$S_\ell(\mathbf{r}_0, \mathbf{r}'_0, \omega) = S_p(\omega)\delta(\mathbf{r}_0 - \mathbf{r}'_p)\delta(\mathbf{r}'_0 - \mathbf{r}_p) \tag{210}$$

The spectrum of the response is, therefore,

$$\begin{aligned} S_w(\mathbf{r}_1, \omega) &= \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} d\mathbf{r}_0\, d\mathbf{r}'_0 G^*(\mathbf{r}_1, \mathbf{r}_0, \omega) \\ &\quad \times G(\mathbf{r}_1, \mathbf{r}'_0, \omega)S_p(\omega)\delta(\mathbf{r}_0 - \mathbf{r}_p)\delta(\mathbf{r}'_0 - \mathbf{r}_p) \qquad (211) \\ &= S_p(\omega)\int_{\mathbf{r}_0} G^*(\mathbf{r}_1, \mathbf{r}_0, \omega) \\ &\quad \times \delta(\mathbf{r}_0 - \mathbf{r}_p)\, d\mathbf{r}_0 \int_{\mathbf{r}'_0} G(\mathbf{r}_1, \mathbf{r}'_0, \omega) \\ &\quad \times \delta(\mathbf{r}'_0 - \mathbf{r}_p)\, d\mathbf{r}'_0 \qquad (212) \\ &= S_p(\omega)|G(\mathbf{r}_1, \mathbf{r}_p, \omega)|^2 \qquad (213) \end{aligned}$$

The spectrum of the response is the square absolute value of the Green's function multiplied by the spectrum of the force. The Green's function in this case is the response at $\mathbf{r}_1$ due to unit sinusoidal force of frequency ω at $\mathbf{r}_p$ (p being the loading point).

Suppose there is a group of *independent* forces on the structure. The cross correlation between them is 0, so:

$$S_\ell(\mathbf{r}_0, \mathbf{r}'_0, \omega) = S(\mathbf{r}_0, \omega)\delta(\mathbf{r}_0 - \mathbf{r}'_0) \tag{214}$$

That is, $S_\ell = 0$ except when $\mathbf{r}_0 = \mathbf{r}'_0$, so:

$$\begin{aligned} S_w(\mathbf{r}_1, \omega) &= \int_{\mathbf{r}_0}\int_{\mathbf{r}'_0} G^*(\mathbf{r}_1, \mathbf{r}_0, \omega)G(\mathbf{r}_1, \mathbf{r}'_0, \omega) \\ &\quad \times S(\mathbf{r}_0, \omega)\delta(\mathbf{r}_0 - \mathbf{r}'_0)\, d\mathbf{r}_0\, d\mathbf{r}'_0 \\ &= \int_{\mathbf{r}'_0} G(\mathbf{r}_1, \mathbf{r}'_0, \omega)\Big[\int_{\mathbf{r}_0} G^*(\mathbf{r}_1, \mathbf{r}_0, \omega) \\ &\quad \times S(\mathbf{r}_0, \omega)\delta(\mathbf{r}_0 - \mathbf{r}'_0)\, d\mathbf{r}_0\Big]\, d\mathbf{r}'_0 \\ &= \int_{\mathbf{r}'_0} G(\mathbf{r}_1, \mathbf{r}'_0, \omega)G^*(\mathbf{r}_1, \mathbf{r}'_0, \omega)S(\mathbf{r}'_0, \omega)]\, d\mathbf{r}'_0 \\ &= \int_{\mathbf{r}'_0} |G(\mathbf{r}_1, \mathbf{r}'_0, \omega)|^2 S(\mathbf{r}'_0, \omega)\, d\mathbf{r}'_0 \qquad (215) \end{aligned}$$

If there are n forces, each with spectrum $S(\mathbf{r}_n, \omega)$,

$$S_w(\mathbf{r}_1, \omega) = \sum_n |G(\mathbf{r}_1, \mathbf{r}_n, \omega)|^2 S(\mathbf{r}_n, \omega) \tag{215a}$$

The response is just the sum of the spectra for each force acting separately.

2. Computation of the Response in Terms of Modes

The general variational equation of motion for any elastic structure can be written as:

$$\begin{aligned} &\iiint_V [\rho(\ddot{u}\delta u + \ddot{v}\delta v + \ddot{w}\delta w) + \delta W]\, dV \\ &\quad - \iint_S (X_\nu \delta u + Y_\nu \delta \upsilon + Z_\nu \delta w)\, dS = 0 \qquad (216) \end{aligned}$$

where ρ is mass density of body; u, v, w are displacements at any point; $\delta u, \delta v, \delta w$ are variations of the displacements; X_ν, Y_ν, Z_ν are surface forces; ds is the elemental surface area; dV is the elemental volume; and δW is the variation of potential energy. In accordance with Love's analysis, let the displacements in the normal modes be described by:

$$u = u_r\varphi'_r, \qquad v = v_r\varphi'_r, \qquad w = w_r\varphi'_r \tag{217}$$

where $\varphi'_r = A_r \cos p_r t$, p_r being the natural frequency of the rth mode. Now let the forced motion of the system be described by:

$$u = \sum_r u_r\varphi_r, \quad v = \sum_r v_r\varphi_r, \quad w = \sum_r w_r\varphi_r \tag{218}$$

where u_r, v_r, w_r are the mode shapes, and φ_r is a function of time. In accordance with Love, let

$$\begin{aligned} u &= u_r\varphi_r \qquad \delta u = u_s\varphi_s \\ v &= v_r\varphi_r \qquad \delta v = v_s\varphi_s \qquad (219) \\ w &= w_r\varphi_r \qquad \delta w = w_s\varphi_s \end{aligned}$$

Substituting into the variational equation of motion, we obtain the following:

$$\begin{aligned} &\iiint_V \rho(u_r\ddot{\varphi}_r u_s\varphi_s + v_r\ddot{\varphi}_r v_s\varphi_s + w_r\ddot{\varphi}_r w_s\varphi_s)\, dV \\ &\quad + \iiint_V \delta W\, dV = \iint_S (X_\nu u_s\varphi_s + Y_\nu v_s\varphi_s \\ &\quad + Z_\nu w_s\varphi_s)\, dS \qquad (220) \end{aligned}$$

However, since the modal functions satisfy the equation for free vibration:

$$\begin{aligned} \iiint_V \delta W\, dV = \iiint_V \rho\big(&p_r^2 u_r\varphi_r u_s\varphi_s \\ &+ p_r^2 v_r\varphi_r v_s\varphi_s + p_r^2 w_r\varphi_r w_s\varphi_s\big)\, dV \end{aligned} \tag{221}$$

and Love shows that:

$$\iiint_V \rho(u_r u_s + v_r v_s + w_r w_s)\, dV = 0 \qquad r \neq s \quad (222)$$

the final equation of motion becomes:

$$\ddot{\varphi}_r(t) + p_r^2 \varphi(t) = F_r(t) \quad (223)$$

where $M_r = \iiint_V \rho(u_r^2 + v_r^2 + w_r^2)\, dV$ (the generalized mass for the rth mode):

$$F_r(t) = \frac{1}{M_r} \iint_S [X_v(t)u_r + Y_v(t)v_r + Z_v(t)w_r]\, dS \quad (224)$$

If structural damping is taken into account, it can be written as another generalized force that opposes the motion:

$$(F_r)_{\text{damping}} = -\kappa \ddot{\varphi}_r \iiint_V \left(u_r^2 + v_r^2 + w_r^2\right) dV \quad (225)$$

where κ is the damping force per unit volume per unit velocity. Finally, the equation of motion becomes:

$$\ddot{\varphi}_r + \psi_r \ddot{\varphi}_r + p_r^2 \varphi_r = F_r \quad (226)$$

where:

$$\psi_r = \frac{\kappa}{M_r} \iiint_V \left(u_r^2 + v_r^2 + w_r^2\right) dV$$

It is convenient to employ the vector notation; thus, let the displacement functions in the rth mode be written as:

$$\mathbf{q}_r = u_r \mathbf{i} + v_r \mathbf{j} + w_r \mathbf{k} \quad (227)$$

where $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are the unit vectors in the x, y, z directions, respectively. Let

$$\mathbf{F}(s,t) = X_v \mathbf{i} + Y_v \mathbf{j} + Z_v \mathbf{k} \quad (228)$$

Thus,

$$\begin{aligned} M_r &= \iiint_V \rho \mathbf{q}_r \cdot \mathbf{q}_r\, dV, \qquad \mathbf{q}_r = \mathbf{q}_r(V) \\ F_r(t) &= \frac{1}{M_r} \iint_S \mathbf{F} \cdot \mathbf{q}_r\, ds, \qquad \mathbf{F} = \mathbf{F}(S,t) \end{aligned} \quad (229)$$

The Fourier transform of $\mathbf{F}(S,t)$ is

$$S_{\mathbf{F}}(S,\omega) = \int_{-\infty}^{+\infty} \mathbf{F}(S,t) e^{-i\omega r}\, dt \quad (230)$$

and the Fourier transform of φ_r is

$$S_{\varphi_r}(\omega) = \int_{-\infty}^{+\infty} \varphi_r(t) e^{-i\omega r}\, dt \quad (231)$$

Now,

$$\begin{aligned} \varphi_r(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_{\varphi_r}(\omega) e^{i\omega t}\, d\omega \\ \dot{\varphi}_r(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} i\omega S_{\varphi_r}(\omega)\, e^{i\omega t}\, d\omega \\ \ddot{\varphi}_r(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} -\omega^2 S_{\varphi_r}(\omega) e^{i\omega t}\, d\omega \end{aligned} \quad (232)$$

$$S_{\dot{\varphi}_r}(\omega) = i\omega S_{\varphi_r}(\omega), \qquad S_{\ddot{\varphi}_r}(\omega) = -\omega^2 S_{\varphi_r}(\omega) \quad (233)$$

Let β_r be the damping constant for the rth mode. Now take the Fourier transform of the equation of motion:

$$S_{\ddot{\varphi}_r} + \beta_r S_{\dot{\varphi}_r} + p_r^2 S_{\varphi_r} = S_{F_r} \quad (234)$$

which is

$$-\omega^2 S_{\varphi_r}(\omega) + i\omega\beta_r S_{\varphi_r}(\omega) + p_r^2 S_{\varphi_r}(\omega) = S_{F_r}(\omega) \quad (235)$$

where:

$$S_{F_r}(\omega) = \frac{1}{M_r} \iint_S S_{\mathbf{F}}(\mathbf{r}_s, \omega) \cdot \mathbf{q}_r(\mathbf{r}_s)\, ds \quad (236)$$

So,

$$S_{\varphi_r}(\omega) = \frac{\iint_s S_{\mathbf{F}}(\mathbf{r}_s, \omega) \cdot \mathbf{q}_r(\mathbf{r}_s)\, ds}{M_r\left[\left(p_r^2 - \omega^2\right) + i\omega\beta_r\right]}$$

In dealing with statistical averaging, the cross correlation function is used. The cross correlation between the displacement at two points in any direction (the direction can be different at the two points) is

$$\Gamma_q(\mathbf{r}_1, \mathbf{r}_2, \tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} q(\mathbf{r}_1, t) q(\mathbf{r}_2, t+\tau)\, dt \quad (237)$$

We are picking a given direction at each point, so the two quantities are scalar (no longer vector). Then,

$$\begin{aligned} q &= \sum_r q_r \varphi_r \\ q(\mathbf{r}_2, t+\tau) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_q(\mathbf{r}_2, \omega) e^{i\omega(t+\tau)}\, d\omega \\ q(\mathbf{r}_1, t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_q(\mathbf{r}_1, \omega) e^{i\omega t}\, d\omega \end{aligned} \quad (238)$$

So,

$$\begin{aligned}\Gamma_q(\mathbf{r}_1, \mathbf{r}_2, \tau) &= \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{+T} q(\mathbf{r}_1, t) \\ &\quad \times \left(\frac{1}{2\pi} \int_{-T}^{+T} S_q(\mathbf{r}_2, \omega) e^{i\omega(t+\tau)}\, d\omega \right) dt \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} S_q(\mathbf{r}_2, \omega) e^{i\omega\tau} \\ &\quad \times \left(\frac{1}{2\pi} \int_{-T}^{+T} q(\mathbf{r}_1, t) e^{i\omega t}\, dt \right) d\omega \\ &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} S_q^T(\mathbf{r}_2, \omega) S_q^*(\mathbf{r}_1, \omega) e^{i\omega\tau}\, d\omega \end{aligned} \tag{239}$$

Now, the power spectral density of the displacement is defined in terms of the cross correlation as:

$$\Gamma_q(\mathbf{r}_1, \mathbf{r}_2, \tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} G_q(\mathbf{r}_1, \mathbf{r}_2, \omega) e^{i\omega\tau}\, d\omega \tag{240}$$

Then,

$$G_q(\mathbf{r}_1, \mathbf{r}_2, \omega) = \lim_{T\to\infty} \frac{1}{2T} S_q^{*T}(\mathbf{r}_1, \omega) S_q^{*T}(\mathbf{r}_2, \omega) \tag{241}$$

Now,

$$S_q^T(\mathbf{r}_2, \omega) = \sum_r q_r(\mathbf{r}_2) S_{\varphi_r}^T(\omega) \tag{242}$$

Thus,

$$\begin{aligned} G_q(\mathbf{r}_1, \mathbf{r}_2, \omega) &= \lim_{T\to\infty} \frac{1}{2T} \sum_r \sum_k \\ &\quad \times q_r(\mathbf{r}_1) S_{\varphi_r}^T(\omega) q_k(\mathbf{r}_2) S_{\varphi_k}^T(\omega) \end{aligned} \tag{243}$$

$$\begin{aligned} G_q(\mathbf{r}_1, \mathbf{r}_2, \omega) &= \sum_r \sum_k \frac{q_r(\mathbf{r}_1) q_k(\mathbf{r}_2)}{Y_r^*(i\omega) Y_k(I\omega)} \int_{S_u} \int_{S_v} \\ &\quad \times \lim_{T\to\infty} \frac{1}{2T} \left[S_F^{*T}\left(\mathbf{r}_{S_u}, \omega\right) \cdot \mathbf{q}_r\left(\mathbf{r}_{S_u}\right) \right] \\ &\quad \times \left[S_F^T\left(\mathbf{r}_{S_v}, \omega\right) \cdot \mathbf{q}_k\left(\mathbf{r}_{S_v}\right) \right] ds_u\, ds_v \end{aligned} \tag{244}$$

Now, if the integrand is written in the double surface integral, it is

$$\begin{aligned} &S_X^{*T} S_X^T u_r u_k + S_X^{*T} S_Y^T u_r v_k + S_X^{*T} S_Z^T u_r w_k + S_Y^{*T} S_X^T v_r u_k \\ &\quad + S_Y^{*T} S_Y^T v_r v_k + S_Y^{*T} S_Z^T v_r w_k + S_Z^{*T} S_X w_r u_k \\ &\quad + S_Z^{*T} S_Y^T w_r v_k + S_Z^{*T} S_Z^T w_r w_k \end{aligned} \tag{245}$$

Note the tensor properties of the last expression involving each component of loading. Note that in the general formula involving the dot product, the component of the modal vector in the direction of the loading function at the two points has to be taken. Now, assuming that the loading is normal to the surface of the structure, our concern is with the cross spectral density of the normal acceleration at two points (or cross spectral density between normal acceleration at two points r_{1s} and r_{2s}) on the surface:

$$\begin{aligned} a_n(\mathbf{r}_{1s}, \mathbf{r}_{2s}, \omega) &= \omega^4 \sum_r \sum_q \frac{[q_r(\mathbf{r}_{1s}) q_k(\mathbf{r}_{2s})]_n}{Y_r^*(i\omega) Y_k(i\omega)} \\ &\quad \times \int_{S_u} \int_{S_v} \int \int G_p\left(\mathbf{r}_{S_u}, \mathbf{r}_{S_v}, \omega\right) \\ &\quad \times q_r\left(\mathbf{r}_{S_u}\right) q_k\left(\mathbf{r}_{S_v}\right) ds_u\, ds_v \end{aligned} \tag{246}$$

where $G_p(\mathbf{r}_{S_u}, \mathbf{r}_{S_v}, \omega)$ is the cross spectral density of the loading normal to the surface at points $\mathbf{r}_{S_u}$ and $\mathbf{r}_{S_v}$. The mean square acceleration over a frequency band Ω_1 to Ω_2 at point $\mathbf{r}_{1S}$ is given by

$$\overline{a_n(\mathbf{r}_{1s})^2_{\Delta\Omega}} = \frac{1}{2\pi} \int_{\Omega_1}^{\Omega_2} a_n(\mathbf{r}_{1S}, \mathbf{r}_{1s}, \omega)\, d\omega \tag{247}$$

Equation (246) is nothing other than Eq. (203) with the integrand expanded in terms of modes of the structure, and Eq. (203) in turn is nothing other than Eq. (128) written for a continuous structure instead of just a linear black box system.

D. Coupled Structural Acoustic Systems

Equation (171) stated that the cross spectral density of the field pressure in tems of acceleration spectra on the surface is

$$\begin{aligned} G(P_1, P_2, \omega) &= \frac{1}{(4\pi)^2} \int_{S_1} \int_{S_2} \rho_0^2 a_n(S_1, S_2, \omega) \bar{g}(P_1, S_1, \omega) \bar{g}^* \\ &\quad \times (P_2, S_2, \omega)\, dS_1\, dS_2 \end{aligned} \tag{248}$$

Furthermore, it was found in the last section that the cross-spectral density of the normal acceleration for a structure in which the loading is normal to the surface can then be written:

$$a_n(S_1, S_2, \omega) = \omega^4 \sum_r \sum_m \frac{q_{rn}(S_1) q_{mn}(S_2)}{Y_r(i\omega) Y_m^*(i\omega)} C_{rm}(\omega) \tag{249}$$

in which:

$$C_{rm}(\omega) = \int_{S_1} \int_{S_2} G(S_1, S_2, \omega) q_{rn}(S_1) q_{mn}(S_2)\, dS_1\, dS_2$$

where $G(S_1, S_2, \omega)$ is the cross-spectral density of the pressure that excites the structure, and $q_{rn}(S_1)$ is the

normal component of the rth mode evaluated at point S_1 of the surface. If the damping in the structure is relatively low, then in accordance with the analysis of Powell, Hurty, and Rubenstein, the cross-product terms can be neglected and

$$a_n(S_1, S_2, \omega) \approx \omega^4 \sum_r \frac{q_{rn}(S_1) q_{rn}(S_2)}{|Y_r(i\omega)|^2} C_{rr}(\omega) \qquad (250)$$

where:

$$C_{rr}(\omega) = \int_{S_1} \int_{S_2} G(S_1, S_2, \omega) q_{rn}(S_1) q_{rn}(S_2)\, dS_1\, dS_2$$

To carry the analysis further, a Green's function must be obtained. Using the analysis of Strasberg and Morse and Ingard as a guide, we assume the use of a free-field Green's function. The analysis, although approximate, then comes out in general form instead of being limited to a particular surface. Therefore, let

$$\bar{g}(P_1, S_1, \omega) = \frac{e^{ikR_1}}{R_1} e^{-ik(\mathbf{a}_{R_1} \cdot \mathbf{R}_{S_1})}\, dS_2 \qquad (251)$$

in which (see Fig. 11):

$$\mathbf{a}_{R_1} \cdot \mathbf{R}_{S_1} = z_o \cos\theta_1 + x_o \sin\theta_1 \cos\varphi_1 + y_o \sin\theta_1 \sin\varphi_1$$

where x_o, y_o, z_o are the rectangular coordinates of the point on the vibrating surface of the structure; $\mathbf{R}_{S_1}$ is the radius vector to point S_1 on the surface; R_1, θ_1, φ_1 are the spherical coordinates of point P_1 in the far field; $\mathbf{a}_{R_1}$ is a unit vector in the direction of $\mathbf{R}_1$ (the radius vector from the origin to the far-field point). Thus, $\mathbf{a}_{R_1} \cdot \mathbf{R}_{S_1}$ is the projection of $\mathbf{R}_{S_1}$ on $\mathbf{R}_1$, *making* $\mathbf{R}_1 - \mathbf{a}_R \cdot \mathbf{R}_{S_1}$ the distance from the far-field point to the surface point. Therefore,

$$\bar{g}^*(P_2, S_2, \omega) = \frac{e^{-ikR_2}}{R_2} e^{ik(\mathbf{a}_{R_2} \cdot \mathbf{R}_{S_2})} \qquad (253)$$

Combining Eqs. (248), (249), (251), and (253) gives the following expression for far-field cross spectrum of far-field pressure:

$$G(P_1, P_2, \omega) = \omega^4 \rho_o^2 \frac{e^{ik(R_1 - R_2)}}{(4\pi)^2 R_1 R_2} \sum_r \sum_m I_r(\theta_1, \varphi_1, \omega) I_m^* \times (\theta_2, \varphi_2, \omega) \times \frac{C_{rm}}{Y_r(i\omega) Y_m^*(i\omega)} \qquad (254)$$

where:

$$I_r = \int_{S_1} q_{rn}(S_1) e^{-ik(\mathbf{a}_{R_1} \cdot \mathbf{R}_{S_1})}\, dS_1$$
$$I_m^* = \int_{S_2} q_{mn}(S_2) e^{-ik(\mathbf{a}_{R_2} \cdot \mathbf{R}_{S_2})}\, dS_2 \qquad (255)$$

With the low damping approximation given by Eq. (250), the far-field auto spectrum at point P_1 is

$$G(P_1, P_1, \omega) \approx \frac{\rho_o \omega^4}{(4\pi)^2 R_1^2} \times \sum_r \frac{|I_r(\theta_1, \varphi_1, \omega)|^2}{|Y_r(i\omega)|^2} C_{rr}(\omega) \qquad (256)$$

The far-field mean square pressure in a frequency band $\Delta\Omega = \Omega_2 - \Omega$ can be written:

$$\overline{p(P_1)}^2_{\Delta\Omega} = \frac{1}{2\pi} \int_{\Omega_1}^{\Omega_2} G(P_1, P_2, \omega)\, dw \qquad (257)$$

Thus,

$$\overline{p(P_1)}^2_{\Delta\Omega} = \frac{\rho_o^2}{(4\pi)^2 R_1^2} \frac{1}{2\pi} \times \int_{\Omega_1}^{\Omega_2} \left[\sum_r \frac{\omega^4 C_{rr}(\omega)}{|Y_r(i\omega)|^2} |I_r(\theta_1, \varphi_1, \omega)|^2\, dw \right] \qquad (258)$$

In cases in which the structure is lightly damped, the following can be written:

$$\frac{1}{2\pi} \int_{\Omega_1}^{\Omega_2} \frac{\omega^4 C_{rr}(\omega)\, d\omega}{|Y_r(i\omega)|^2} \approx \frac{p_r C_{rr}(p_r)}{8\zeta_r M_r^2} \qquad (259)$$

Where $C_{rr}(p_r)$ is defined as the joint acceptance evaluated at the natural frequency p_r (p_r consists of those natural frequencies between Ω_1 and Ω_2), M_r is the total generalized mass of the rth mode (including virtual mass), and

$$\zeta_r = \bar{C}_r / (\bar{C}_c)_r \qquad (260)$$

where $\bar{C}_r$ is the damping constant for the rth mode (including radiation damping) and $(\bar{C}_c)_r$ is the critical damping constant for that mode. Thus, the mean square pressure at the far-field point P_1 in the frequency band $\Delta\Omega$ is

$$\overline{p(P_1)}^2_{\Delta\Omega} \approx \frac{\rho_o^2}{R_1^2} \frac{1}{(4\pi)^2} \sum_{r\ in\ \Delta\Omega} \frac{p_r C_{rr}(p_r)}{8\zeta_r M_r^2} |I_r(\theta_1, \varphi_1, p_r)|^2 \qquad (261)$$

In Eq. (261), $C_{rr}(p_r)$ describes the characteristics of the generalized force of the random loading, $p_r / 8\zeta_r M_r^2$ describes the characteristics of the structure, and I_r describes the directivity of the noise field. The sum is taken over those modes that resonate in the band.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • ACOUSTIC CHAOS • ACOUSTIC WAVE DEVICES • SIGNAL PROCESSING, ACOUSTIC • WAVE PHENOMENA

BIBLIOGRAPHY

Ando, Y. (1998). "Architectural Acoustics: Blending Sound Sources, Sound Fields, and Listener," *Modern Acoustics and Signal Processing*, Springer-Verlag.

Brekhovskikh, L. M., and Godin, O. A. (1998). "Acoustics of Layered Media I: Plane and Quasi-Plane Waves," *Springer Series on Wave Phenomena,* Vol. 5, Springer-Verlag.

Brekhovskikh, L. M., and Godin, O. A. (1999). "Acoustics of Layered Media II: Point Sources and Bounded Beams," *Springer Series on Wave Phenomena,* Vol. 14, Springer-Verlag.

Howe, M. S. (1998). "Acoustics of Fluid-Structure Interactions," Cambridge University Press.

Kishi, T., Ohtsu, M., and Yuyama, S., eds. (2000). "Acoustic Emission—Beyond the Millennium," Elsevier.

Munk, W., Worcester, P., and Wunsch, C. (1995). "Ocean Acoustic Tomography," Cambridge University Press.

Ohayon, R., and Soize, C. (1998). "Structural Acoustics and Vibration," Academic Press.

Tohyama, M., Suzuki, H., and Ando, Y. (1996). "The Nature and Technology of Acoustic Space," Academic Press, London.



Chaos

Joshua Socolar
Duke University

I. Introduction
II. Classical Chaos
III. Dissipative Dynamical Systems
IV. Hamiltonian Systems
V. Quantum Chaos

GLOSSARY

Cantor set Simple example of a fractal set of points with a noninteger dimension.

Chaos Technical term referring to the irregular, unpredictable, and apparently random behavior of deterministic dynamical systems.

Deterministic equations Equations of motion with no random elements for which formal existence and uniqueness theorems guarantee that once the necessary initial and boundary conditions are specified the solutions in the past and future are uniquely determined.

Dissipative system Dynamical system in which frictional or dissipative effects cause volumes in the phase space to contract and the long-time motion to approach an attractor consisting of a fixed point, a periodic cycle, or a strange attractor.

Dynamical system System of equations describing the time evolution of one or more dependent variables. The equations of motion may be difference equations if the time is measured in discrete units, a set of ordinary differential equations, or a set of partial differential equations.

Ergodic theory Branch of mathematics that introduces statistical concepts to describe average properties of deterministic dynamical systems.

Extreme sensitivity to initial conditions Refers to the rapid, exponential divergence of nearby trajectories in chaotic dynamical systems.

Fractal Geometrical structure with self-similar structure on all scales that may have a noninteger dimension, such as the outline of a cloud, a coastline, or a snowflake.

Hamiltonian system Dynamical system that conserves volumes in phase space, such as a mechanical oscillator moving without friction, the motion of a planet, or a particle in an accelerator.

KAM theorem The Kolmogorov-Arnold–Moser theorem proves that when a small, nonlinear perturbation is applied to an integrable Hamiltonian system it remains nearly integrable if the perturbation is sufficiently small.

Kicked rotor Simple model of a Hamiltonian dynamical system that is exactly described by the classical standard map and the quantum standard map.

Kolmogorov–Sinai entropy Measure of the rate of mixing in a chaotic dynamical system that is closely related

to the average Lyapunov exponent, which measures the exponential rate of divergence of nearby trajectories.

Localization Quantum interference effect, introduced by Anderson in solid-state physics, which inhibits the transport of electrons in disordered or chaotic dynamical systems as in the conduction of electronics in disordered media or the microwave excitation and ionization of highly excited hydrogen atoms.

Lyapunov exponent A real number λ specifying the average exponential rate at which nearby trajectories in phase space diverge or converge.

Mixing Technical term from ergodic theory that refers to dynamical behavior that resembles the evolution of cream poured in a stirred cup of coffee.

Period-doubling bifurcations Refers to a common route from regularity to chaos in chaotic dynamical systems in which a sequence of periodic cycles appears in which the period increases by a factor of two as a control parameter is varied.

Phase space Mathematical space spanned by the dependent variables of the dynamical system. For example, a mechanical oscillator moving in one dimension has a two-dimensional phase space spanned by the position and momentum variables.

Poincaré section Stroboscopic picture of the evolution of a dynamical system in which the values of two dependent variables are plotted as points in a plane each time the other dependent variables assume a specified set of values.

Random matrix theory Theory introduced to describe the statistical fluctuations of the spacings of nuclear energy levels based on the statistical properties of the eigenvalues of matrices with random elements.

Resonance overlap criterion Simple analytical estimate of the conditions for breakup of KAM surfaces leading to widespread, global chaos.

Strange attractor Aperiodic attracting set with a fractal structure that often characterizes the longtime dynamics of chaotic dissipative systems.

Trajectory A path in the phase space of a dynamical system that is traced out by a system starting from a particular set of initial values of the dependent variables.

Universality Refers to the detailed quantitative similarity of the transition from regular behavior to chaos in a broad class of disparate dynamical systems.

A WIDE VARIETY of natural phenomena exhibit complex, irregular behavior. In the past, many of these phenomena were considered to be too difficult to analyze; however, the advent of high-speed digital computers coupled with new mathematical and physical insight has led to the development of a new interdisciplinary field of science called *nonlinear dynamics*, which has been very successful in finding some underlying order concealed in nature's complexity. In particular, research in the latter half on the 20th century has revealed how very simple, diterministic mathematical models of physical and biological systems can exhibit surprisingly complex behavior. The apparently random behavior of these deterministic, nonlinear dynamical systems is called *chaos*.

Since many different fields of science and engineering are confronted with difficult problems involving nonlinear equations, the field of nonlinear dynamics has evolved in a highly interdisciplinary manner, with important contributions coming from biologists, mathematicians, engineers, and physicists. In the physical sciences, important advances have been made in our understanding of complex processes and patterns in dissipative systems, such as damped, driven, nonlinear oscillators and turbulent fluids, and in the derivation of statistical descriptions of Hamiltonian systems, such as the motion of celestial bodies and the motion of charged particles in accelerators and plasmas. Moreover, the predictions of chaotic behavior in simple mechanical systems have led to the investigation of the manifestations of chaos in the corresponding quantum systems, such as atoms and molecules in very strong fields.

This article attempts to describe some of the fundamental ideas; to highlight a few of the important advances in the study of chaos in classical, dissipative, and Hamiltonian systems; and to indicate some of the implications for quantum systems.

I. INTRODUCTION

In the last 25 years, the word *chaos* has emerged as a technical term to refer to the complex, irregular, and apparently random behavior of a wide variety of physical phenomena, such as turbulent fluid flow, oscillating chemical reactions, vibrating structures, the behavior of nonlinear electrical circuits, the motion of charged particles in accelerators and fusion devices, the orbits of asteroids, and the dynamics of atoms and molecules in strong fields. In the past, these complex phenomena were often referred to as random or stochastic, which meant that researchers gave up all hope of providing a detailed microscopic description of these phenomena and restricted themselves to statistical descriptions alone. What distinguishes chaos from these older terms is the recognition that many complex physcial phenomena are actually described by deterministic equations, such as the Navier–Stokes equations of fluid mechanics, Newton's equations of classical mechanics, or Schrödinger's equation of quantum mechanics, and the important discovery that even very simple, deterministic equations of motion can exhibit exceedingly

complex behavior and structure that is indistinguishable from an idealized random process. Consequently, a new term was required to describe the irregular behavior of these deterministic dynamical systems that reflected the new found hope for a deeper understanding of these various physical phenomena. These realizations also led to the rapid development of a new, highly interdisciplinary field of scientific research called *nonlinear dynamics*, which is devoted to the description of complex, but deterministic, behavior and to the search for "order in chaos."

The rise of nonlinear dynamics was stimulated by the combination of some old and often obscure mathematics from the early part of the 20th century that were preserved and developed by isolated mathematicians in the United States, the Soviet Union, and Europe; the deep natural insight of a number of pioneering researchers in meteorology, biology, and physics; and by the widespread availability of high-speed digital computers with high-resolution computer graphics. The mathematicians constructed simple, but abstract, dynamical systems that could generate complex behavior and geometrical patterns. Then, early researchers studying the nonlinear evolution of weather patterns and the fluctuations of biological populations realized that their crude approximations to the full mathematical equations, in the form of a single difference equation or a few ordinary differential equations, could also exhibit behavior as complex and seemingly random as the natural phenomena. Finally, high-speed computers provided a means for detailed computer experiments on these simple mathematical models with complex behavior. In particular, high-resolution computer graphics have enabled experimental mathematicians to search for order in chaos that would otherwise be buried in reams of computer output. This rich interplay of mathematical theory, physical insight, and computer experimentation, which characterizes the study of chaos and the field of nonlinear dynamics, will be clearly illustrated in each of the examples discussed in this article.

Chaos research in the physical sciences and engineering can be divided into three distinct areas relating to the study of nonlinear dynamical systems that correspond to (1) classical dissipative systems, such as turbulent flows or mechanical, electrical, and chemical oscillators; (2) classical Hamiltonian systems, where dissipative processes can be neglected, such as charged particles in accelerators and magnetic confinement fusion devices or the orbits of asteroids and planets; and (3) quantum systems, such as atoms and molecules in strong static or intense electromagnetic fields or electrons confined to submicron-scale cavities.

The study of chaos in classical systems (both dissipative and Hamiltonian) is now a fairly well-developed field that has been described in great detail in a number of popular and technical books. In particular, the term *chaos* has a very precise mathematical definition for classical nonlinear systems, and many of the characteristic features of chaotic motion, such as the extreme sensitivity to initial conditions, the appearance of strange attractors with noninteger fractal dimensions, and the period-doubling route to chaos, have been cataloged in a large number of examples and applications, and new discoveries continue to fill technical journals. In Section II, we will begin with a precise definition of chaos for classical systems and present a very simple mathematical example that illustrates the origin of this complex, apparently random motion in simple deterministic dynamical systems. In Section II, we will also consider additional examples to illustrate some of the other important general features of chaotic classical systems, such as the notion of geometric structures with noninteger dimensions.

Some of the principal accomplishments of the application of these new ideas to dissipative systems include the discovery of a universal theory for the transition from regular, periodic behavior to chaos via a sequence of period-doubling bifurcations, which provides quantitative predictions for a wide variety of physical systems, and the discoveries that mathematical models of turbulence with as few as three nonlinear differential equations can exhibit chaotic behavior that is governed by a strange attractor.

The ideas and analytical methods introduced by the simple models of nonlinear dynamics have provided important analogies and metaphors for describing complex natural phenomena that should ultimately pave the way for a better theoretical understanding. Section III will be devoted to a detailed discussion of several models of dissipative systems with important applications in the description of turbulence and the onset of chaotic behavior in a variety of nonlinear oscillators.

The latter portion of Section III introduces concepts associated with the description of dissipative dynamical systems with many degrees of freedom and briefly discusses some issues that have been central to chaos research in the last decade.

In the realm of Hamiltonian systems, the exact nonlinear equations for the motion of particles in accelerators and fusion devices and of celestial bodies are simple enough to be analyzed using the analytical and numerical methods of nonlinear dynamics without any gross approximations. Consequently, accurate quantitative predictions of the conditions for the onset of chaotic behavior that play significant roles in the design of accelerators and fusion devices and in understanding the irregular dynamics of asteroids can be made. Moreover, the important realization that only a few interacting particles, representing a small number of degrees of freedom, can exhibit motion that is

sufficiently chaotic to permit a statistical description has greatly enhanced our understanding of the microscopic foundations of statistical mechanics, which have also remained an outstanding problem of theoretical physics for over a century. Section IV will examine several simple mathematical models of Hamiltonian systems with applications to the motion of particles in accelerators and fusion devices and to the motion of celestial bodies.

Finally, in Section V, we will discuss the more recent and more controversial studies of the quantum behavior of strongly coupled and strongly perturbed Hamiltonian systems, which are classically chaotic. In contrast to the theory of classical chaos, there is not yet a consensus on the definition of quantum chaos because the Schrodinger equation is a linear equation for the deterministic evolution of the quantum wave function, which is incapable of exhibiting the strong dynamical instability that defines chaos in nonlinear classical systems. Nevertheless, both numerical studies of model problems and real experiments on atoms and molecules reveal that quantum systems can exhibit behavior that resembles classical chaos for long times. In addition, considerable research has been devoted to identifying the distinct signatures or symptoms of the quantum behavior of classically chaotic systems. At present, the principal contributions of these studies has been the demonstration that atomic and molecular physics of strongly perturbed and strongly coupled systems can be very different from that predicted by the traditional perturbative methods of quantum mechanics. For example, experiments with highly excited hydrogen atoms in strong microwave fields have revealed a novel ionization mechanism that depends strongly on the intensity of the radiation but only weakly on the frequency. This dependence is just the opposite of the quantum photoelectric effect, but the sharp onset of ionization in the experiments is very well described by the onset of chaos in the corresponding classical system.

II. CLASSICAL CHAOS

This section provides a summary of the fundamental ideas that underlie the discussion of chaos in all classical dynamical systems. It begins with a precise definition of chaos and illustrates the important features of the definition using some very simple mathematical models. These examples are also used to exhibit some important properties of chaotic dynamical systems, such as extreme sensitivity to initial conditions, the unpredictability of the long-time dynamics, and the possibility of geometric structures corresponding to strange attractors with noninteger, fractal dimensions. The manifestations of these fundamental concepts in more realistic examples of dissipative and Hamiltonian systems will be provided in Sections III and IV.

A. The Definition of Chaos

The word *chaos* describes the irregular, unpredictable, and apparently random behavior of nonlinear dynamical systems that are described mathematically by the deterministic iteration of nonlinear difference equations or the evolution of systems of nonlinear ordinary or partial differential equations. The precise mathematical definition of chaos requires that the dynamical system exhibit mixing behavior with positive Kolmogorov–Sinai entropy (or positive average Lyapunov exponent). This definition of chaos invokes a number of concepts from ergodic theory, which is a branch of mathematics that arose in response to attempts to reconcile statistical mechanics with the deterministic equations of classical mechanics. Although the equations that describe the evolution of chaotic dynamical systems are fully deterministic (no averages over random forces or initial conditions are involved), the complexity of the dynamics invites a statistical description. Consequently, the statistical concepts of ergodic theory provide a natural language to define and characterize chaotic behavior.

1. Ergodicity

A central concept familiar to physicists because of its importance to the foundations of statistical mechanics the notion of ergodicity. Roughly speaking, a dynamical system is ergodic if the system comes arbitrarily close to every possible point (or state) in the accessible phase space over time. In this case, the celebrated ergodic theorem guarantees that long-time averages of any physical quantity can be determined by performing averages over phase space with respect to a probability distribution. However, although there has been considerable confusion in the physical literature, ergodicity alone is not sufficiently irregular to account for the complex behavior of turbulent flows or interacting many-body systems. A simple mathematical example clearly reveals these limitations.

Consider the dynamical system described by the difference equation,

$$x_{n+1} = x_n + a, \qquad \text{Mod } 1 \tag{1}$$

which takes a real number x_n between 0 and 1, adds another real number a, and subtracts the integer part of the sum (Mod 1) to return a value of x_{n+1} on the unit interval [0, 1]. The sequence of numbers, $\{x_n\}_{n=0,1,2,3,\ldots}$, generated by iterating this one-dimensional map describes the time history of the dynamical variable x_n (where time is measured in discrete units labeled by n). If $a = p/q$ is a rational number (where p and q are integers), then starting with any initial x_0, this dynamical system generates a time

sequence of $\{x_n\}$ that returns to x_0 after q iterations since $x_q = x_0 + p \text{ (Mod1)} = x_0 \text{ (Mod 1)}$. In this case, the long-time behavior is described by a periodic cycle of period q that visits only q different values of x on the unit interval [0, 1]. Since this time sequence does not come arbitrarily close to every point in the unit interval (which is the phase or state space of this dynamical system), this map is not ergodic for rational values of a.

However, if a is an irrational number, the time sequence never repeats and x_n will come arbitrarily close to every point in the unit interval. Moreover, since the time sequence visits every region of the unit interval with equal probability, the long-time averages of any functions of the dynamical variable x can be replaced by spatial averages with respect to the uniform probability distribution $P(x) = 1$ for x in [0, 1]. Therefore, for irrational values of a, this dynamical system, described by a single, deterministic difference equation, is an ergodic system.

Unfortunately, the time sequence generated by this map is much too regular to be chaotic. For example, if we initially colored all the points in the phase space between 0 and $\frac{1}{4}$ red and iterated the map, then the red points would remain clumped together in a continuous interval (Mod 1) for all time. But, if we pour a little cream in a stirred cup of coffee or release a dyed gas in the corner of the room, the different particles of the cream or the colored gas quickly spread uniformly over the accessible phase space.

2. Mixing

A stronger notion of statistical behavior is required to describe turbulent flows and the approach to equilibrium in many-body systems. In ergodic theory, this property is naturally called *mixing*. Roughly speaking, a dynamical system described by deterministic difference or differential equations is said to be a mixing system if sets of initial conditions that cover limited regions of the phase space spread throughout the accessible phase space and evolve in time like the particles of cream in coffee. Once again a simple difference equations serves to illustrate this concept.

Consider the shift map:

$$x_{n+1} = 2x_n, \qquad \text{Mod } 1 \tag{2}$$

which takes x_n on the unit interval, multiplies it by 2, and subtracts the integer part to return a value of x_{n+1} on the unit interval. If we take almost any initial condition, x_0, then this deterministic map generates a time sequence $\{x_n\}$ that never repeats and for long times is indistinguishable from a random process. Since the successive iterates wander over the entire unit interval and come arbitrarily close to every point in the phase space, this map is ergodic. Moreover, like Eq. (1), the long-time averages of any function of the $\{x_n\}$ can be replaced by the spatial average with respect to the uniform probability distribution $P(x) = 1$.

However, the dynamics of each individual trajectory is much more irregular than that generated by Eq. (1). If we were to start with a set of red initial conditions on the interval $[0, \frac{1}{4}]$, then it is easy to see that these points would be uniformly dispersed on the unit interval after only two iterations of the map. Therefore, we call this dynamical system a *mixing system*. (Of course, if we were to choose very special initial conditions, such as $x_0 = 0$ or $x_0 = p/2^m$, where p and m are positive integers, then the time sequence would still be periodic. However, in the set of all possible initial conditions, these exceptional initial conditions are very rare. Mathematically, they comprise a set of zero measure, which means the chance of choosing one of these special initial conditions by accident is nil.)

It is very easy to see that the time sequences generated by the vast majority of possible initial conditions is as random as the time sequence generated by flipping a coin. Simply write the initial condition in binary representation, that is, $x_0 = 0.0110011011100011010\ldots$. Multiplication by 2 corresponds to a register shift that moves the binary point to the right (just like multiplying a decimal number by 10). Therefore, when we iterate Eq. (2), we read off successive digits in the initial condition. If the leading digit to the left of the binary point is a one, then the Mod 1 replaces it by a 0. Since a theorem by Martin–Löf guarantees that the binary digits of almost every real number are a random sequence with no apparent order, the time sequence $\{x_n\}$ generated by iterating this map will also be random. In particular, if we call out heads whenever the leading digit is a 1 (which means that x_n lies on the interval $[\frac{1}{2}, 1]$) and tails whenever the leading digit is a 0 (which means that x_n lies on the interval $[0, \frac{1}{2}]$), then the time sequence $\{x_n\}$ generated by this deterministic difference equation will jump back and forth between the left and right halves of the unit interval in a process that is indistinguishable from that generated by a series of coin flips.

The technical definition of chaos refers to the behavior of the time sequence generated by a mixing system, such as the shift map defined by Eq. (2). This simple, deterministic dynamical system with random behavior is the simplest chaotic system, and it serves as the paradigm for all chaotic systems.

3. Extreme Sensitivity to Initial Conditions

One of the essential characteristics of chaotic systems is that they exhibit extreme sensitivity to initial conditions. This means that two trajectories with initial conditions that

are arbitrarily close will diverge at an exponential rate. The exponential rate of divergence in mixing systems is related to a positive Kolmogorov–Sinai entropy. For simple systems, such as the one-dimensional maps defined by Eqs. (1) and (2), this local instability is characterized by the average Lyapunov exponent, which in practice is much easier to evaluate than the Kolmogorov–Sinai entropy.

It is easy to see that Eq. (2) exhibits extreme sensitivity to initial conditions with a positive average Lyapunov exponent, while Eq.(1) does not. If we consider two nearby initial conditions x_0 and y_0, which are $d_0 = |x_0 - y_0|$ apart, then after one iteration of a map, $x_{n+1} = F(x_n)$ of the form of Eqs. (1) or (2), the two trajectories will be approximately separated by a distance $d_1 = |(dF/dx)(x_0|d_0)$. Clearly, if $|dF/dx| < 1$, the distance between the two points decreases; if $|dF/dx| > 1$, the two distance increases; while if $|dF/dx| = 1$, the two trajectories remain approximately the same distance apart. We can easily see by differentiating the map or looking at the slopes of the graphs of the return maps in Figs. 1 and 2 that $|dF/dx| = 1$ for Eq. (1), while $|dF/dx| = 2$ for Eq. (2). Therefore, after many iterations of Eq. (1), nearby initial conditions will generate trajectories that stay close together (the red points remained clumped), while the trajectories generated by Eq. (2) diverge at an exponential rate (the red points mix throughout the phase space). Moreover, the average Lyapunov exponent for these one-dimensional maps, defined by:

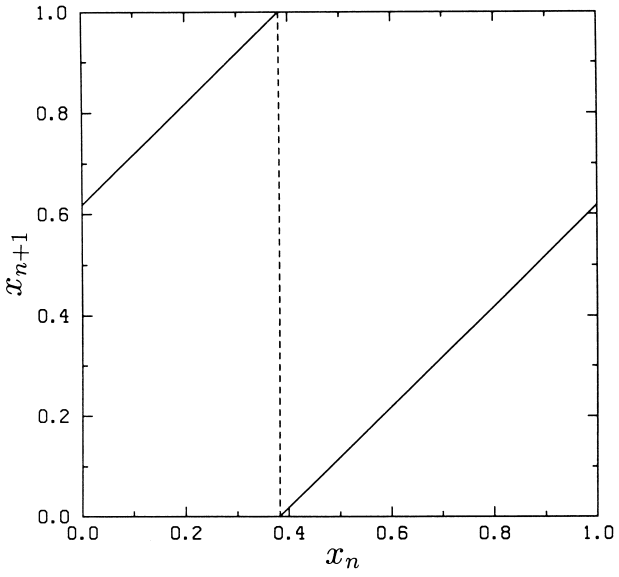


FIGURE 1 A graph of the return map defined by Eq. (1) for $a = (\sqrt{5}-1)/2 \simeq 0.618$. The successive values of the time sequence $\{x_n\}_{n=1,2,3,\ldots}$ are simply determined by taking the old values of x_n and reading off the new values x_{n+1} from the graph.

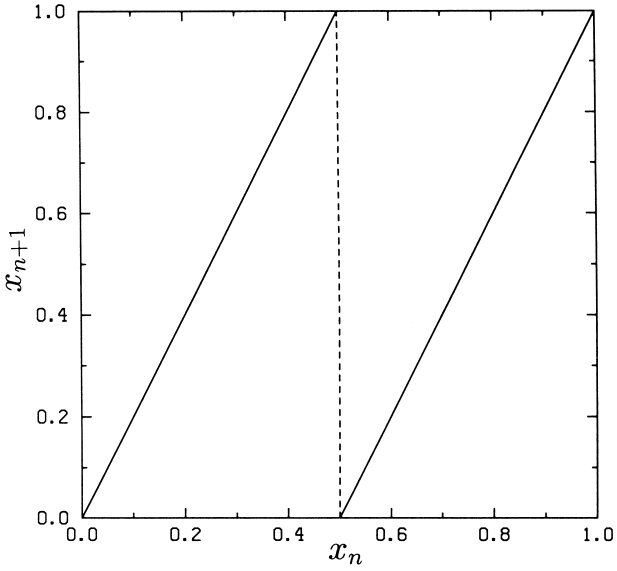


FIGURE 2 A graph of the return map defined by Eq. (2). For values of x_n between 0 and 0.5, the map increases linearly with slope 2; but for x_n larger than 0.5, the action of the Mod 1 requires that the line reenter the unit square at 0.5 and rise again to 1.0.

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} \ln \left| \frac{dF}{dx}(x_n) \right| \tag{3}$$

provides a direct measure of the exponential rate of divergence of nearby trajectories. Since the slope of the return maps for Eqs. (1) and (2) are the same for almost all values of x_n, the average Lyapunov exponents can be easily evaluated. For Eq. (1), we get $\lambda = 0$, while Eq. (2) gives $\lambda = \log 2 > 0$.

However, it is important to note that all trajectories generated by Eq. (2) do not diverge exponentially. As mentioned earlier, the set of rational x_0's with even denominators generate regular periodic orbits. Although these points are a set of measure zero compared with all of the real numbers on the unit interval, they are dense, which means that in every subinterval, no matter how small, we can always find one of these periodic orbits. The significance of these special trajectories is that, figuratively speaking, they play the role of rocks or obstructions in a rushing stream around which the other trajectories must wander. If this dense set of periodic points were not present in the phase space, then the extreme sensitivity to initial conditions alone would not be sufficient to guarantee mixing behavior. For example, if we iterated Eq. (2) without the Mod 1, then all trajectories would diverge exponentially, but the red points would never be able to spread throughout the accessible space that would consist of the entire positive real axis. In this case, the dynamical system is simply unstable, not chaotic.

4. Unpredictability

One important consequence of the extreme sensitivity to initial conditions is that long-term prediction of the evolution of chaotic dynamical systems, like predicting the weather, is a practical impossibility. Although chaotic dynamical systems are fully deterministic, which means that once the initial conditions are specified the solution of the differential or difference equations are uniquely determined for all time, this does not mean that it is humanly possible to find the solution for all time. If nearby initial conditions diverge exponentially, then any errors in specifying the initial conditions, no matter how small, will also grow exponentially. For example, if we can only specify the initial condition in Eq. (2) to an accuracy of one part in a thousand, then the uncertainty in predicting x_n will double each time step. After only 10 time steps, the uncertainty will be as large as the entire phase space, so that even approximate predictions will be impossible. (If we can specify the initial conditions to double precision accuracy on a digital computer (1 part in 10^{18}), then we can only provide approximate predictions of the future values of x_n for 60 time steps before the error spans the entire unit interval.) In contrast, if we specify the initial condition for Eq. (1) to an accuracy of 10^{-3}, then we can always predict the future values of x_n to the same accuracy. (Of course, errors in the time evolution can also arise from uncertainties in the parameters in the equations of evolution. However, if we can only specify the parameter a in Eq. (1) to an accuracy of 10^{-3}, we could still make approximate predictions for the time sequence for as many as 10^3 iterations before the uncertainty becomes as large as the unit interval.)

B. Fractals

Another common feature of chaotic dynamical systems is the natural appearance of geometrical structures with noninteger dimensions. For example, in dissipative dynamical systems, described by systems of differential equations, the presence of dissipation (friction) causes the long-time behavior to converge to a geometrical structure in the phase space called an *attractor*. The attractor may consist of a single fixed point with dimension 0, a periodic limit cycle described by a closed curve with dimension 1, or, if the long-time dynamics is chaotic, the attracting set may resemble a curve with an infinite number of twists, turns, and folds that never closes on itself. This strange attractor is more than a simple curve with dimension 1, but it may fail to completely cover an area of dimension 2. In addition, strange attractors are found to exhibit the same level of structure on all scales. If we look at the complex structure through a microscope, it does not look any simpler no matter how much we increase the magnifications. The term *fractal* was coined by Benoit Mandelbrot to describe these complex geometrical objects. Like the shapes of snowflakes and clouds and the outlines of coastlines and mountain ranges, these fractal objects are best characterized by a noninteger dimension.

The simplest geometrical object with a noninteger dimension is the middle-thirds Cantor set. If we take the unit interval [0, 1] and remove all of the points in the middle third, then we will be left with a set consisting of two pieces $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$ each of length $\frac{1}{3}$. If we remove the middle thirds of these remaining pieces. we get a set consisting of four pieces of length $\frac{1}{9}$. By repeating this construction *ad infinitum*, we end up with a strange set of points called a Cantor set. Although it consists of points none is isolated. In fact, if we magnify any interval containing elements of the set—for example, the segment contained on the interval $[0, \frac{1}{3^n}]$ for any positive n—then the magnified interval will look the same as the complete set (see Fig. 3).

In order to calculate a dimension for this set, we must first provide a mathematical definition of dimension that agrees with our natural intuition for geometrical objects with integer dimension. Although there are a variety of

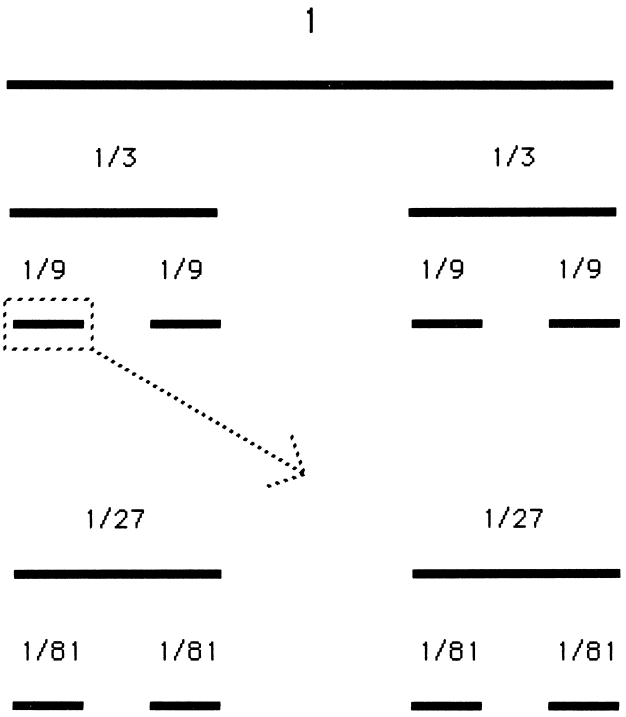


FIGURE 3 The middle-thirds Cantor set is constructed by first removing the points in the middle third of the unit interval and then successively removing the middle thirds of the remaining intervals *ad infinitum*. This figure shows the first four stages of Cantor set construction. After the first two steps, the first segment is magnified to illustrate the self-similar structure of the set.

different definitions of dimension corresponding to different levels of mathematical rigor, one definition that serves our purpose is to define the dimension of a geometrical object in terms of the number of boxes of uniform size required to cover the object. For example, if we consider two-dimensional boxes with sides of length L (for example, $L = 1$ cm), then the number of boxes required to cover a two-dimensional object, $N(L)$, will be approximately equal to the area measured in units of L^2 (that is, cm^2). Now, if we decrease the size of the boxes to L', then the number of boxes $N(L')$ will increase approximately as $(L/L'^2$. [If $L' = 1$ mm, then $N(L') \approx 100N(L)$.] Similarly, if we try to cover a one-dimensional object, such as a closed curve, with these boxes, the number of boxes will increase only as (L/L'). (For $L = 1$ cm and $L' = 1$ mm, $N(L)$ will be approximately equal to the length of the curve in centimeters, while $N(L')$ will be approximately the length of the curve in millimeters.) In general,

$$N(L) \propto L^{+d} \tag{4}$$

where d is the dimension of the object. Therefore, one natural mathematical definition of dimension is provided by the equation:

$$d = \lim_{L \to 0} \log N(L) / \log(1/L) \tag{5}$$

obtained by taking the logarithm of both sides of Eq. (4). For common geometrical objects, such as a point, a simple curve, an area, or a volume, this definition yields the usual integer dimensions 0, 1, 2, and 3, respectively. However, for the strange fractal sets associated with many chaotic dynamical systems, this definition allows for the possibility of noninteger values. For example, if we count the number of boxes required to cover the middle-thirds Cantor set at each level of construction, we find that we can always cover every point in the set using 2^n boxes of length $\left(\frac{1}{3}\right)^n$ (that is, 2 boxes of length $\frac{1}{3}$, 4 boxes of length $\frac{1}{9}$, etc.), Therefore, Eq. (5) yields a dimension of $d = \log 2 / \log 3 = 0.63093\ldots$, which reflects the intricate self-similar structure of this set.

III. DISSIPATIVE DYNAMICAL SYSTEMS

In this section, we will examine three important examples of simple mathematical models that can exhibit chaotic behavior and that arise in applications to problems in science and engineering. Each represents a dynamical system with dissipation so that the long-time behavior converges to an attractor in the phase space. The examples increase in complexity from a single difference equation, such as Eqs. (1) and (2), to a system of two coupled difference equations and then to a system of three coupled ordinary differential equations. Each example illustrates the characteristic properties of chaos in dissipative dynamical systems with irregular, unpredictable behavior that exhibits extreme sensitivity to initial conditions and fractal attractors.

A. The Logistic Map

The first example is a one-dimensional difference equation, like Eqs. (1) and (2), called the logistic map, which is defined by:

$$X_{n+1} = a x_n (1 - x_n) \equiv F(X_n) \tag{6}$$

For values of the control parameter a between 0 and 4, this nonlinear difference equation also takes values of x_n between 0 and 1 and returns a value x_{n+1} on the unit interval. However, as a is varied, the time sequences $\{x_n\}$ generated by this map exhibit extraordinary transitions from regular behavior, such as that generated by Eq. (1), to chaos, such as that generated by Eq. (2). Although this mathematical model is too simple to be directly applicable to problems in physics and engineering, which are usually described by differential equations, Mitchell Feigenbaum has shown that the transition from order to chaos in dissipative dynamical systems exhibits universal characteristics (to be discussed later), so that the logistic map is representative of a large class of dissipative dynamical systems. Moreover, since the analysis of this deceptively simple difference equation involves a number of standard techniques used in the study of nonlinear dynamical systems, we will examine it in considerable detail.

As noted in the seminal review article in 1974 by Robert May, a biologist who considered the logistic map as a model for annual variations of insect populations, the time evolution generated by the map can be easily studied using a graphical analysis of the return maps displayed in Fig. 4. Equation (6) describes an inverted parabola that intercepts the $x_{n+1} = 0$ axis at $x_n = 0$ and 1, with a maximum of $x_{n+1} = a/4$ at $x_n = 0.5$. Although this map can be easily iterated using a short computer program, the qualitative behavior of the time sequence $\{x_n\}$ generated by any initial x_0 can be examined by simply tracing lines on the graph of the return map with a pencil as illustrated in Fig. 4.

For values of $a < 1$, almost every initial condition is attracted to $x = 0$ as shown in Fig. 4 for $a = 0.95$. Clearly, $x = 0$ is a fixed point of the nonlinear map. If we start with $x_0 = 0$, then the logistic map returns the value $x_n = 0$ for all future iterations. Moreover, a simple linear analysis, such as that used to define the Lyapunov exponent in Section II, shows that for $a < 1$ this fixed point is stable. (Initial conditions that are slightly displaced from the origin will be attracted back since $|(dF/dx)(0)| = a < 1$.)

However, when the control parameter is increased to $a > 1$, this fixed point becomes unstable and the long-time behavior is attracted to a new fixed point, as shown in Fig. 4 for $a = 2.9$, which lies at the other intersection of

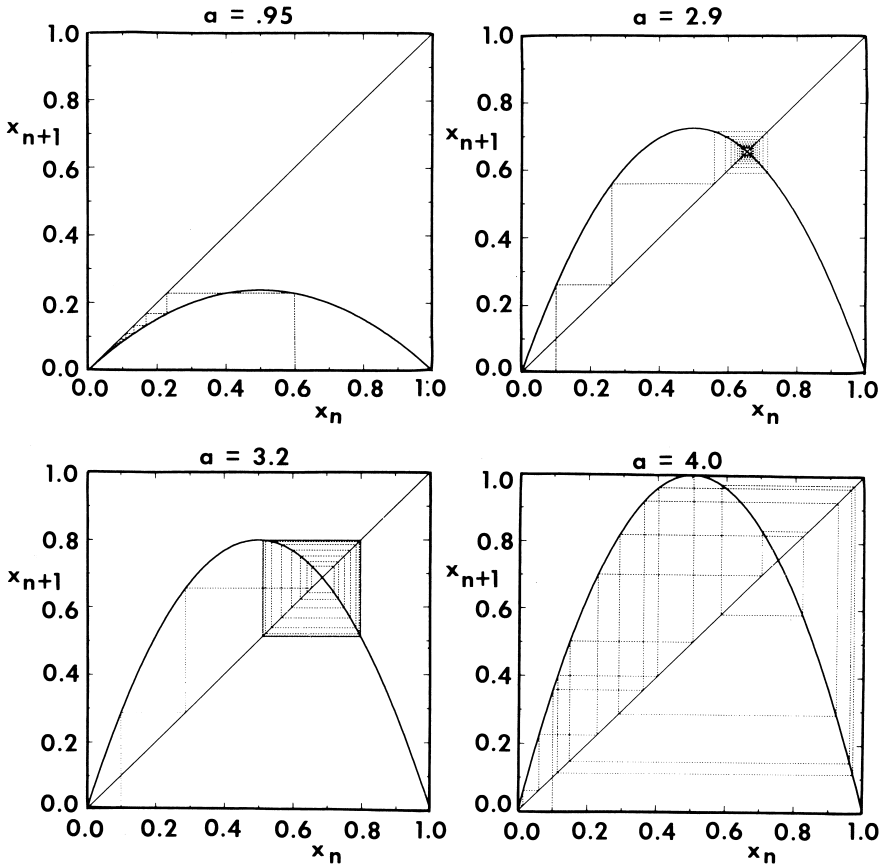


FIGURE 4 Return maps for the logistic map, Eq. (6), are shown for four different values of the control parameter a. These figures illustrate how pencil and paper can be used to compute the time evolution of the map. For example, if we start our pencil at an initial value of $x_0 = 0.6$ for $a = 0.95$, then the new value of x_1 is determined by tracing vertically to the graph of the inverted parabola. Then, to get x_2, we could return to the horizontal axis and repeat this procedure, but it is easier to simply reflect off the 45° line and return to the parabola. Successive iterations of this procedure give rapid convergence to the stable, fixed point at $x = 0$. However, if we start at $x_0 = 0.1$ for $a = 2.9$, our pencil computer diverges from $x = 0$ and eventually settles down to a stable, fixed point at the intersection of the parabola and the 45° line. Then, when we increase $a > 3$, this fixed point repels the trace of the trajectory, which settles into either a periodic cycle, such as the period-2 cycle for $a = 3.2$, or a chaotic orbit, such as that for $a = 4.0$.

the 45° line and the graph of the return map. In this case, the dynamical system approaches an equilibrium with a nonzero value of the dependent variable x. Elementary algebra shows that this point corresponds to the nonzero root of the quadratic equation $x = ax(1 - x)$ given by $x^* = (a - 1)/a$. Again, a simple linear analysis of small displacements from this fixed point reveals that it remains stable for values of a between 1 and 3. When a becomes larger than 3, this fixed point also becomes unstable and the long-time behavior becomes more complicated, as shown in Fig. 4.

1. Period Doubling

For values of a slightly bigger than 3, empirical observations of the time sequences for this nonlinear dynamical system generated by using a hand calculator, a digital computer, or our "pencil computer" reveals that the long-time behavior approaches a periodic cycle of period 2, which alternates between two different values of x. Because of the large nonlinearity in the difference equation, this periodic behavior could not be deduced from any analytical arguments based on exact solutions or from perturbation theory. However, as typically occurs in the field of nonlinear dynamics, the empirical observations provide us with clues to new analytical procedures for describing and understanding the dynamics. Once again, the graphical analysis provides an easy way of understanding the origin of the period-2 cycle.

Consider a new map.

$$\begin{aligned} x_{n+2} &= F^{(2)}(x_n) = F[F(x_n)] \\ &= a^2\left(x_n - x_n^2\right) - a^3\left(x_n^2 - 2x_n^3 + x_n^4\right) \end{aligned} \tag{7}$$

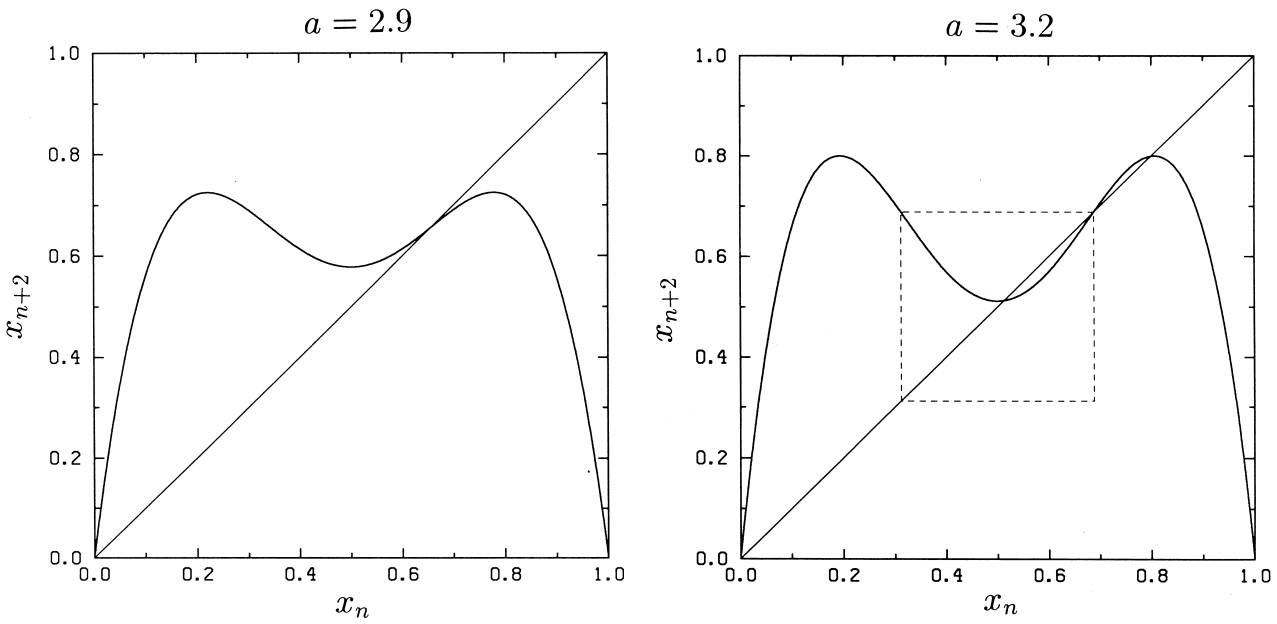


FIGURE 5 The return maps are shown for the second iterate of the logistic map, $F^{(2)}$, defined by Eq. (7). The fixed points at the intersection of the 45° line and the map correspond to values of x that repeat every two periods. For $a = 2.9$, the two intersections are just the period-1 fixed points at 0 and a^*, which repeat every period and therefore every other period, as well. However, when a is increased to 3.2, the peaks and valleys of the return map become more pronounced and pass through the 45° line and two new fixed points appear. Both of the old, fixed points are now unstable because the absolute value of the slope of the return map is larger than 1, but the new points are stable, and they correspond to the two elements of the period-2 cycle displayed in Fig. 4. Moreover, because the portion of the return map contained in the dashed box resembles an inverted image of the original logistic map, one might expect that the same bifurcation process will be repeated for each of these period-2 points as a is increased further.

constructed by composing the logistic map with itself. The graph of the corresponding return map, which gives the values of x_n every other iteration of the logistic map, is displayed in Fig. 5. If we use the same methods of analysis as we applied to Eq. (6), we find that there can be at most four fixed points that correspond to the intersection of the graph of the quartic return map with the 45° line. Because the fixed points of Eq. (4) are values of x that return every other iteration, these points must be members of the period-2 cycles of the original logistic map. However, since the period-1 fixed points of the logistic map at $x = 0$ and x^* are automatically period-2 points, two of the fixed points of Eq. (7) must be $x = 0, x^*$. When $1 < a < 3$, these are the only two fixed points of Eq. (7), as shown in Fig. 5 for $a = 2.9$. However, when a is increased above 3, two new fixed points of Eq. (7) appear, as shown in Fig. 5 for $a = 3.2$, on either side of the fixed point at $x = x^*$, which has just become unstable.

Therefore, when the stable period-1 point at x^* becomes unstable, it gives birth to a pair of fixed points, $x^{(1)}$, $x^{(2)}$ of Eq. (7), which form the elements of the period-2 cycle found empirically for the logistic map. This process is called a *pitchfork bifurcation*. For values of a just above 3, these new fixed points are stable and the long-time dynamics of the second iterate of the logistic map, $F^{(2)}$, is attracted to one or the other of these fixed points. However, as a increases, the new fixed points move away from x^*, the graphs of the return maps for Eq. (7) get steeper and steeper, and when $|dF^{(2)}/dx|_{x^{(1)},x^{(2)}} > 1$ the period-2 cycle also becomes unstable. (A simple application of the chain rule of differential calculus shows that both periodic points destabilize at the same value of a, since $F(x^{(1),(2)}) = x^{(2),(1)}$ and $(dF^{(2)}/dx)(x^{(1)}) = (dF/dx)(x^{(2)})(dF/dx)(x^{(1)}) = (dF^{(2)}/dx)(x^{(1)})$.)

Once again, empirical observations of the long-time behavior of the iterates of the map reveal that when the period-2 cycle becomes unstable it gives birth to a stable period-4 cycle. Then, as a increases, the period-4 cycle becomes unstable and undergoes a pitchfork bifurcation to a period-16 cycle, then a period-32 cycle, and so on. Since the successive period-doubling bifurcations require smaller and smaller changes in the control parameter, this bifurcation sequence rapidly accumulates to a period cycle of infinite period at $a_\infty = 3.57\ldots$.

This sequence of pitchfork bifurcations is clearly displayed in the bifurcation diagram shown in Fig. 6. This

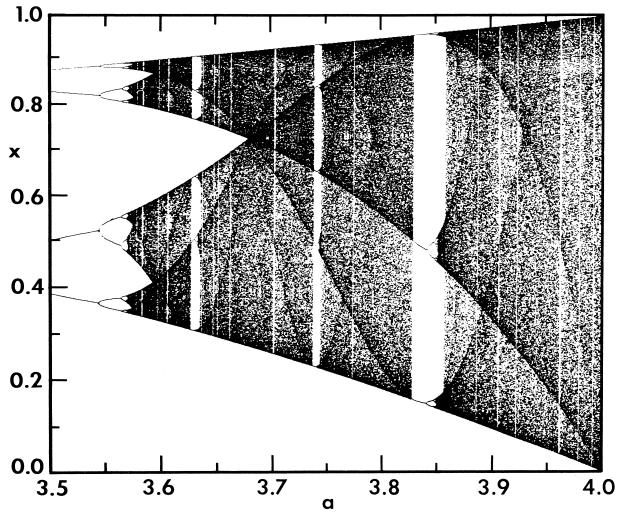


FIGURE 6 A bifurcation diagram illustrates the variety of long-time behavior exhibited by the logistic map as the control parameter a is increased from 3.5 to 4.0. The sequences of period-doubling bifurcations from period-4 to period-8 to period-16 are clearly visible in addition to ranges of a in which the orbits appear to wander over continuous intervals and ranges of a in which periodic orbits, including odd periods, appear to emerge from the chaos.

graph is generated by iterating the map for several hundred time steps for successive values of a. For each value of a, we plot only the last hundred values of x_n to display the long-time behavior. For $a < 3$, all of these points land close to the fixed point at a^*; for $a > 3$, these points alternate between the two period-2 points, then between the four period-4 points, and so on.

The origin of each of these new periodic cycles can be qualitatively understood by applying the same analysis that we used to explain the birth of the period-2 cycle from period 1. For the period-4 cycle, we consider the second iterate of the period-2 map:

$$\begin{aligned} x_{n+4} &= F^{(4)}(x_n) = F^{(2)}\left[F^{(2)}(x_n)\right] \\ &= F\{F[F(F(x_n)]\} \end{aligned} \tag{8}$$

In this case, the return map is described by a polynomial of degree 16 that can have as many as 16 fixed points that correspond to intersections of the 45° line with the graph of the return map. Two of these period-4 points correspond to the period-1 fixed points at 0 and x^*, and for $a > 3$, two correspond to the period-2 points at $x^{(1)}$ and $x^{(2)}$. The remaining 12 period-4 points can form three different period-4 cycles that appear for different values of a. Figure 7 shows a graph of $F^{(4)}(x_n)$ for $a = 3.2$, where the period-2 cycle is still stable, and for $a = 3.5$, where the unstable period-2 cycle has bifurcated into a period-4 cycle. (The other two period-4 cycles are only briefly stable for other values of $a > a_\infty$.)

We could repeat the same arguments to describe the origin of period 8; however, now the graph of the return map of the corresponding polynomial of degree 32 would begin to tax the abilities of our graphics display terminal as well as our eyes. Fortunately, the "slaving" of the stability properties of each periodic point via the chain-rule argument (described previously for the period-2 cycle) means that we only have to focus on the behavior of the successive iterates of the map in the vicinity of the periodic point closest to $x = 0.5$. In fact, a close examination of Figs. 4, 5, and 7 reveals that the bifurcation process for each $F^{(n)}$ is simply a miniature replica of the original period-doubling bifurcation from the period-1 cycle to the period-2 cycle. In each case, the return map is locally described by a parabolic curve (although it is not exactly a parabola beyond the first iteration and the curve is flipped over for every other $F^{(N)}$.

Because each successive period-doubling bifurcation is described by the fixed points of a return map $x_{n+N} = F^{(N)}(X_n)$ with ever greater oscillations on the unit interval, the amount the parameter a must increase before the next bifurcation decreases rapidly, as shown in the bifurcation diagram in Fig. 6. The differences in the changes in the control parameter for each succeeding bifurcation, $a_{n+1} - a_n$, decreases at a geometric rate that is found to rapidly converge to a value of:

$$\delta = \frac{a_n - a_{n-1}}{a_{n+1} - a_n} = 4.6692016\ldots \tag{9}$$

In addition, the maximum separation of the stable daughter cycles of each pitchfork bifurcation also decreases rapidly, as shown in Fig. 6, by a geometric factor that rapidly converges to:

$$\alpha = 2.502907875\ldots \tag{10}$$

2. Universality

The fact that each successive period doubling is controlled by the behavior of the iterates of the map, $F^{(N)}(x)$, near $x = 0.5$, lies at the root of a very significant property of nonlinear dynamical systems that exhibit sequences of period-doubling bifurcations called universality. In the process of developing a quantitative description of period doubling in the logistic map, Feigenbaum discovered that the precise functional form of the map did not seem to matter. For example, he found that a map on the unit interval described by $F(x) = a \sin \pi x$ gave a similar sequence of period-doubling bifurcations. Although the values of the control parameter a at which each period-doubling bifurcation occurs are different, he found that both the ratios of the changes in the control parameter and the separations of the stable daughter cycles decreased at the same geometrical rates δ and α as the logistic map.

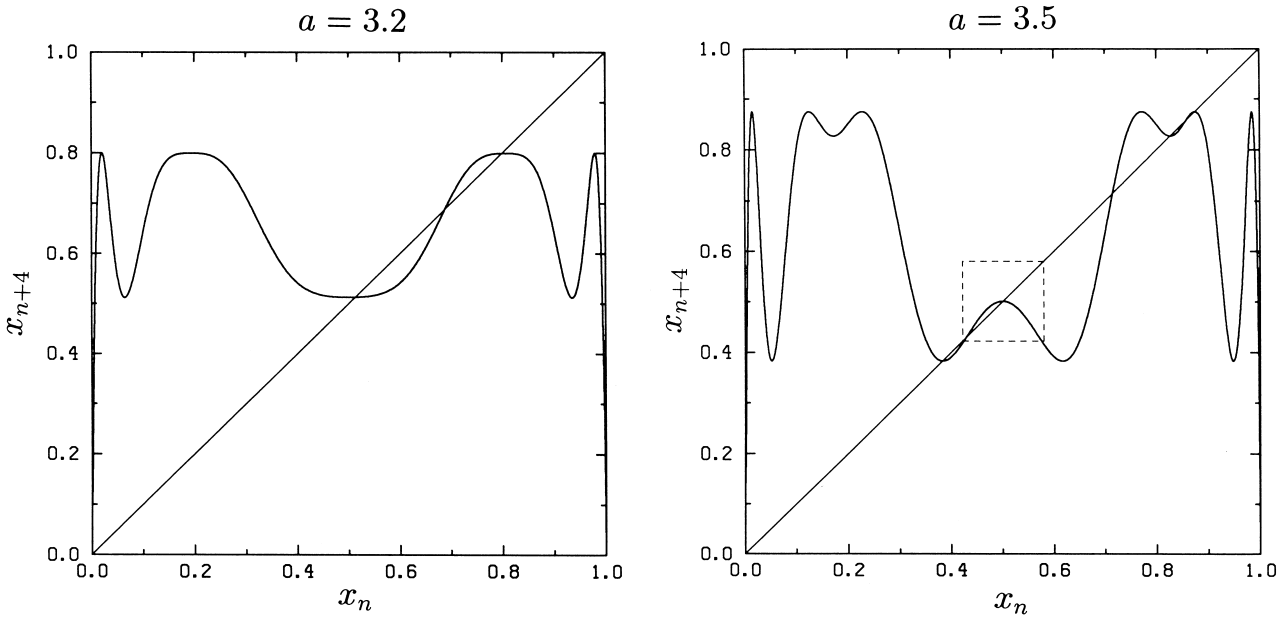


FIGURE 7 The appearance of the period-4 cycle as a is increased from 3.2 to 3.5 is illustrated by these graphs of the return maps for the fourth iterate of the logistic map, $F^{(4)}$. For $a = 3.2$, there are only four period-4 fixed points that correspond to the two unstable period-1 points and the two stable period-2 points. However, when a is increased to 3.5, the same process that led to the birth of the period-2 fixed points is repeated again in miniature. Moreover, the similarity of the portion of the map near $x_n = 0.5$ to the original map indicates how this same bifurcation process occurs again as a is increased.

This observation ultimately led to a rigorous proof, using the mathematical methods of the renormalization group borrowed from the theory of critical phenomena, that these geometrical ratios were universal numbers that would apply to the quantitative description of any period-doubling sequence generated by nonlinear maps with a single quadratic extremum. The logistic map and the sine map are just two examples of this large universality class. The great significance of this result is that the global details of the dynamical system do not matter. A thorough understanding of the simple logistic map is sufficient for describing both qualitatively and, to a large extent, quantitatively the period-doubling route to chaos in a wide variety of nonlinear dynamical systems. In fact, we will see that this universality class extends beyond one-dimensional maps to nonlinear dynamical systems described by more realistic physical models corresponding to two-dimensional maps, systems of ordinary differential equations, and even partial differential equations.

3. Chaos

Of course, these stable periodic cycles, described by Feigenbaum's universal theory, are not chaotic. Even the cycle with an infinite period at the period-doubling accumulation point a has a zero average Lyapunov exponent. However, for many values of a above a_∞, the time sequences generated by the logistic map have a positive average Lyapunov exponent and therefore satisfy the definition of chaos. Figure 8 plots the average Lyapunov exponent computed numerically using Eq. (3) for the same range of values of a, as displayed in the bifurcation diagram in Fig. 6.

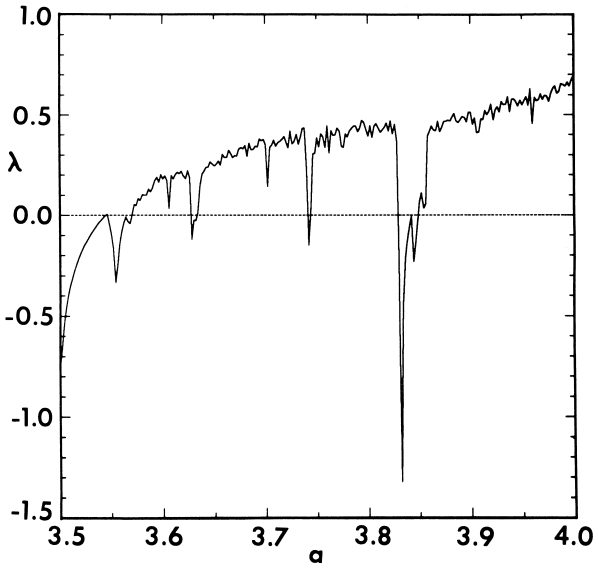


FIGURE 8 The values of the average Lyapunov exponent, computed numerically using Eq. (3), are displayed for the same values of a shown in Fig. 6. Positive values of λ correspond to chaotic dynamics, while negative values represent regular, periodic motion.

Wherever the trajectory appears to wander chaotically over continuous intervals, the average Lyapunov exponent is positive. However, embedded in the chaos for $a_\infty < a < 4$ we see stable period attractors in the bifurcation diagram with sharply negative average Lyapunov exponents. The most prominent periodic cycle is the period-3 cycle, which appears near $a_3 = 3.83$. In fact, between a_∞ and a_3, there is a range of values of a for which cycles of every odd and even period are stable. However, the intervals for the longer cycles are too small to discern in Fig. 6. The period-5 cycle near $a = 3.74$ and the period-6 cycles near $a = 3.63$ and $a = 3.85$ are the most readily apparent in both the bifurcation diagram and the graph of the average Lyapunov exponent.

Although these stable periodic cycles are mathematically dense over this range of control parameters, the values of a where the dynamics are truly chaotic can be mathematically proven to be a significant set with nonzero measure. The proof of the positivity of the average Lyapunov exponent is much more difficult for the logistic map than for Eq. (2) since $\log |(dF/dx)(x_n)|$ can take on both negative and positive values depending on whether x_n is close to $\frac{1}{2}$ or to 0 or 1. However, one simple case for which the logistic map is easily proven to be chaotic is for $a = 4$. In this case, the time sequence appears to wander over the entire unit interval in the bifurcation diagram, and the numerically computed average Lyapunov exponent is positive. If we simply change variables from x_n to $y_n = (2/\pi)\ \sin^{-1}\sqrt{x_n}$, then the logistic map for $a = 4$ transforms to the tent map:

$$y_{n+1} = \begin{cases} 2y_n & 0 \le y_n \le 0.5 \\ 2(1 - y_n) & 0.5 \le y_n \le 1 \end{cases} \tag{11}$$

which is closely related to the shift map, Eq. (2). In particular, since $|dF/dy| = 2$, the average Lyapunov exponent is found to be $\lambda = \log 2 \approx 0.693$, which is the same as the numerical value for the logistic map.

B. The Hénon Map

Most nonlinear dynamical systems that arise in physical applications involve more than one dependent variable. For example, the dynamical description of any mechanical oscillator requires at least two variables—a position and a momentum variable. One of the simplest dissipative dynamical systems that describes the coupled evolution of two variables was introduced by Michel Hénon in 1976. It is defined by taking a one-dimensional quadratic map for x_{n+1} similar to the logistic map and coupling it to a second linear map for y_{n+1}:

$$x_{n+1} = 1 - ax_n^2 + y_n \tag{12a}$$

$$y_{n+1} = bx_n \tag{12b}$$

This pair of difference equations takes points in the x–y plane with coordinates (x_n, y_n) and maps them to new points (x_{n+1}, y_{n+1}). The behavior of the sequence of points generated by successive iterates of this two-dimensional map from an initial point (x_0, y_0) is determined by the values of two control parameters a and b. If a and b are both 0, then Eq. (12) maps every point in the plane to the attracting fixed point at (1, 0) after at most two iterations.

If $b = 0$ but a is nonzero, then the Hénon map reduces to a one-dimensional quadratic map that can be transformed into the logistic map by shifting the variable x. Therefore, for $b = 0$ and even for b small, the behavior of the time sequence of points generated by the Hénon map closely resembles the behavior of the logistic map. For small values of a, the long-time iterates are attracted to stable periodic orbits that exhibit a sequence of period-doubling bifurcations to chaos as the nonlinear control parameter a is increased. For small but nonzero b, the main difference from the one-dimensional maps is that these regular orbits of period N are described by N points in the $(x$–$y)$ plane rather than points on the unit interval. (In addition, the basin of attraction for these periodic cycles consists of a finite region in the plane rather than the unit interval alone. Just as in the one-dimensional logistic map, if a point lies outside this basin of attraction, then the successive iterates diverge to ∞.)

The Hénon map remains a dissipative map with time sequences that converge to a finite attractor as long as b is less than 1. This is easy to understand if we think of the action of the map as a coordinate transformation in the plane from the variables (x_n, y_n) to (x_{n+1}, y_{n+1}). From elementary calculus, we know that the Jacobian of this coordinate transformation, which is given by the determinant of the matrix,

$$M = \begin{pmatrix} -2ax & 1 \\ b & 0 \end{pmatrix} \tag{13}$$

describes how the area covered by any set of points increases or decreases under the coordinate transformation. In this case, $J = \text{Det}\ M = -b$. When $|J| > 1$, areas grow larger and sets of initial conditions disperse throughout the x–y plane under the iteration of the map. But, when $|J| < 1$, the areas decrease under each iteration, so areas must contract to sets of points that correspond to the attractors.

1. Strange Attractors

However, these attracting sets need not be a simple fixed point or a finite number of points forming a periodic cycle. In fact, when the parameters a and b have values that give

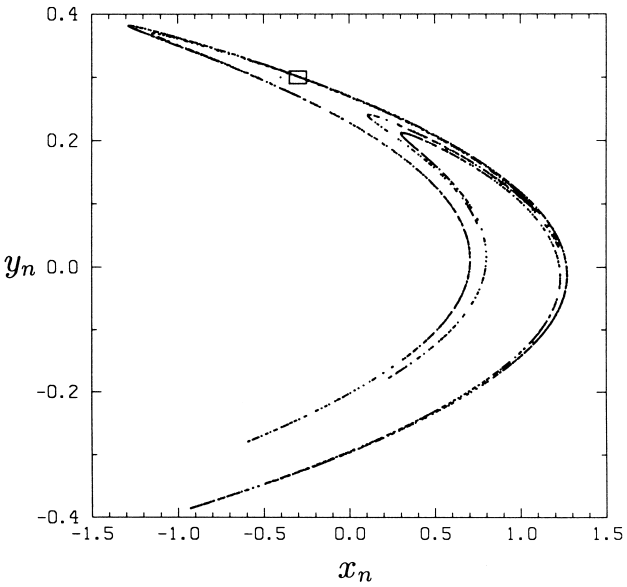


FIGURE 9 The first 10,000 iterates of the two-dimensional Hénon map trace the outlines of a strange attractor in the $x_n - y_n$ plane. The parameters were chosen to be $a = 1.4$ and $b = 0.3$ and the initial point was (0, 0).

rise to chaotic dynamics, the attractors can be exceedingly complex, composed of an uncountable set of points that form intricate patterns in the plane. These strange attractors are best characterized as fractal objects with noninteger dimensions.

Figure 9 displays 10,000 iterates of the Hénon map for $a = 1.4$ and $b = 0.3$. (In this case, the initial point was chosen to be $(x_0, y_0) = (0, 0)$, but any initial point in the basin of attraction would give similar results.) Because $b < 1$, the successive iterates rapidly converge to an intricate geometrical structure that looks like a line that is folded on itself an infinite number of times. The magnifications of the sections of the attractor shown in Fig. 10 display the detailed self-similar structure. The cross sections of the folded line resemble the Cantor set described in Section II.B. Therefore, since the attractor is more than a line but less than an area (since there are always gaps between the strands at every magnification), we might expect it to be characterized by a fractal dimension that lies between 1 and 2. In fact, an application of the box-counting definition of fractal dimension given by Eq. (5) yields a fractal dimension of $d = 1.26\ldots$.

Moreover, if you were to watch a computer screen while these points are plotted, you would see that they wander about the screen in a very irregular manner, slowly revealing this complex structure. Numerical measurements of the sensitivity to initial conditions and of the average Lyapunov exponents (which are more difficult to compute than for one-dimensional maps) indicate that the dynamics on this strange attractor are indeed chaotic.

C. The Lorenz Attractor

The study of chaos is not restricted to nonlinear difference equations such as the logistic map and the Hénon map. Systems of coupled nonlinear differential equations also exhibit the rich variety of behavior that we have already seen in the simplest nonlinear dynamical systems described by maps. A classic example is provided by the Lorenz model described by three coupled nonlinear differential equations:

$$dx/dt = -\sigma x + \sigma y \tag{14a}$$

$$dy/dt = -xz + rx - y \tag{14b}$$

$$dz/dt = xy - bz \tag{14c}$$

These equations were introduced in 1963 by Edward Lorenz, a meteorologist, as a severe truncation of the Navier–Stokes equations describing Rayleigh–Benard convection in a fluid (like Earth's atmosphere), which is heated from below in a gravitational field. The dependent variable x represents a single Fourier mode of the stream function for the velocity flow, the variables y and z represent two Fourier components of the temperature field; and the constants r, σ, and b are the Rayleigh number, the Prandtl number, and a geometrical factor, respectively.

The Lorenz equations provide our first example of a model dynamical system that is reasonably close to a real physical system. (The same equations provide an even better description of optical instabilities in lasers, and similar equations have been introduced to describe chemical oscillators.) Numerical studies of the solutions of these equations, starting with Lorenz's own pioneering work using primitive digital computers in 1963, have revealed the same complexity as the Hénon map. In fact, Hénon originally introduced Eq. (12) as a simple model that exhibits the essential properties of the Lorenz equations.

A linear analysis of the evolution of small volumes in the three-dimensional phase space spanned by the dependent variables x, y, and z shows that this dissipative dynamical system rapidly contracts sets of initial conditions to an attractor. When the Rayleigh number r is less than 1, the point $(x, y, z) = (0, 0, 0)$ is an attracting fixed point. But, when $r > 1$, a wide variety of different attractors that depend in a complicated way on all three parameters r, σ, and b are possible. Like the Hénon map, the long-time behavior of the solutions of these differential equations can be attracted to fixed points; to periodic cycles, which are described by limit cycles consisting of closed curves in the

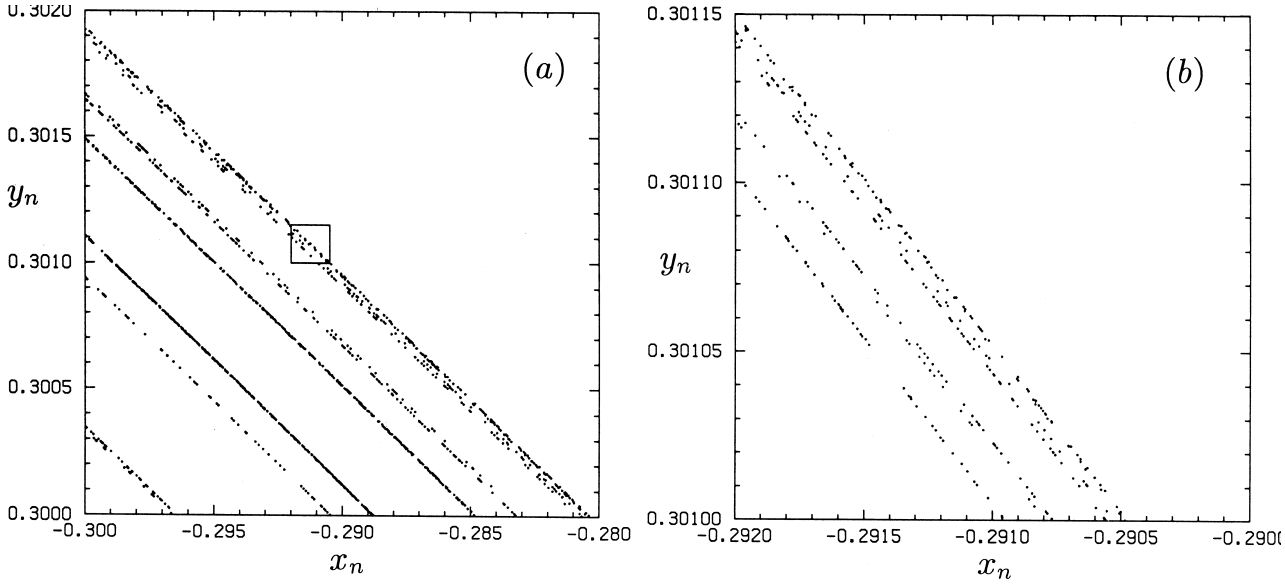


FIGURE 10 To see how strange the attractor displayed in Fig. 9 really is, we show two successive magnifications of a strand of the attractor contained in the box in Fig. 9. Here, (a) shows that the single strand in Fig. 9 breaks up into several distinct bands, which shows even finer structure in (b) when the map is iterated 10,000,000 time steps.

three-dimensional phase space; and to strange attractors, which are described by a fractal structure in phase space. In the first two cases, the dynamics is regular and predictable, but the dynamics on strange attractors is chaotic and unpredictable (as unpredictable as the weather).

The possibility of strange attractors for three or more autonomous differential equations, such as the Lorenz model, was established mathematically by Ruelle and Takens. Figure 11 shows a three-dimensional graph of the famous strange attractor for the Lorenz equations corresponding to the values of the parameters $r = 28$, $\sigma = 10$, and $b = \frac{8}{3}$, which provides a graphic illustration of the consequences of their theorem. The initial conditions were chosen to be $(1, 1, 1)$. The trajectory appears to loop around on two surfaces that resemble the wings of a butterfly, jumping from one wing to the other in an irregular manner. However, a close inspection of these surfaces reveals that under successive magnification they exhibit the same kind of intricate, self-similar structure as the striations of the Hénon attractor. This detailed structure is best revealed by a so-called Poincaré section of continuous dynamics, shown in Fig. 12, which was generated by plotting a point in the x–z plane every time the orbit passes from negative y to positive y.

Since we could imagine that this Poincaré section was generated by iterating a pair of nonlinear difference equations, such as the Hénon map, it is easy to understand, by analogy with the analysis described in Section III.B, how the time evolution can be chaotic with extreme sensitivity to initial conditions and how this cross section of the Lorenz attractor, as well as the Lorenz attractor itself, can have a noninteger, fractal dimension.

D. Applications

Perhaps the most significant conclusion that can be drawn from these three examples of dissipative dynamical systems that exhibit chaotic behavior is that the essential features of the behavior of the more realistic Lorenz model are well described by the properties of the much simpler Hénon map and to a large extent by the logistic map. These observations provide strong motivation to hope that simple nonlinear systems will also capture the essential properties of even more complex dynamical systems that describe a wide variety of physical phenomena with irregular behavior. In fact, the great advances of nonlinear dynamics and the study of chaos in the last [20] years can be attributed to the fulfillment of this hope in both numerical studies of more complicated mathematical models and experimental studies of a variety of complicated natural phenomena.

The successes of this program of reducing the essential features of complicated dynamical processes to simple nonlinear maps or to a few coupled, nonlinear differential equations have been well documented in a number of conference proceedings and textbooks. For example, the universality of the period-doubling route to chaos and the appearance of strange attractors have been demonstrated in numerical studies of a wide variety of nonlinear maps, systems of nonlinear, ordinary, and partial differential equations. Even more importantly, Feigenbaum's

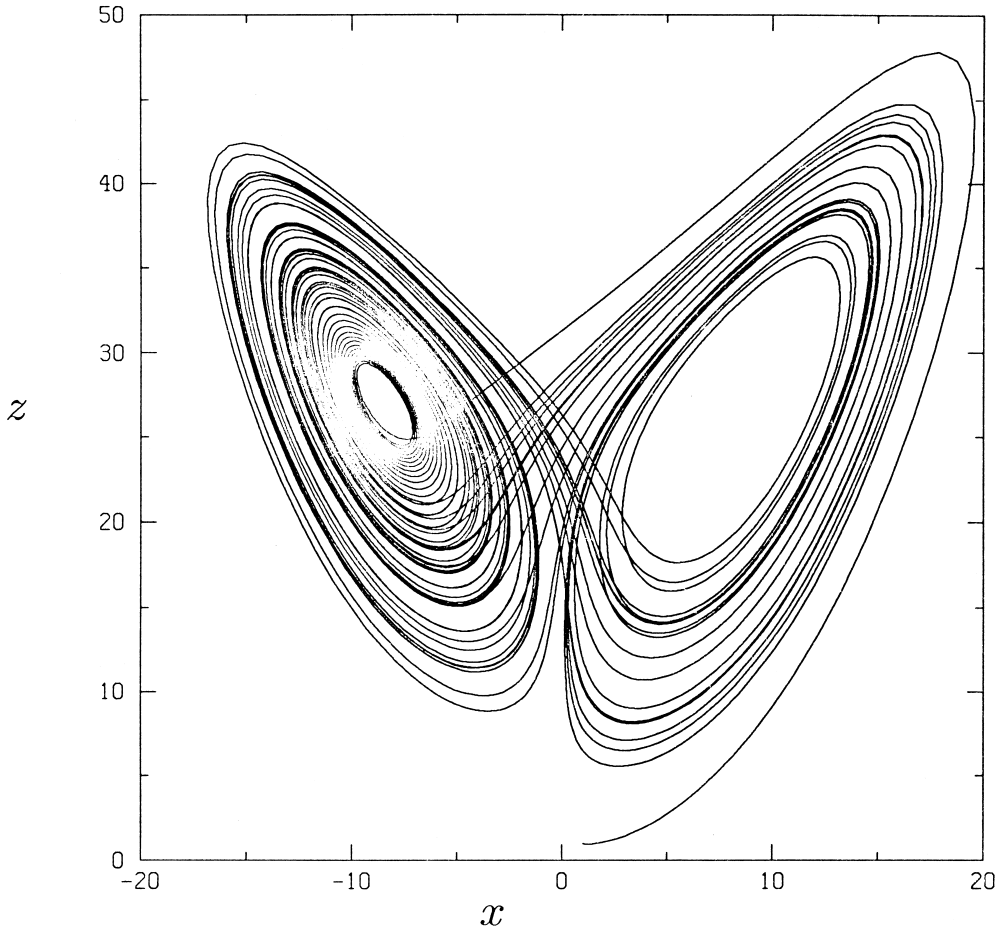


FIGURE 11 The solution of the Lorenz equations for the parameters $r = 28$, $\sigma = 10$, and $b = \frac{8}{3}$ rapidly converges to a strange attractor. This figure shows a projection of this three-dimensional attractor onto the x–z plane, which is traced out by approximately 100 turns of the orbit.

universal constants δ and α, which characterize the quantitative scaling properties of the period-doubling sequence, have been measured to good accuracy in a number of careful experiments on Rayleigh–Benard convection and nonlinear electrical circuits and in oscillating chemical reactions (such as the Belousov–Zabotinsky reaction), laser oscillators, acoustical oscillators, and even the response of heart cells to electrical stimuli. In addition, a large number of papers have been devoted to the measurement of the fractal dimensions of strange attractors that may govern the irregular, chaotic behavior of chemical reactions, turbulent flows, climatic changes, and brainwave patterns.

Perhaps the most important lesson of nonlinear dynamics has been the realization that complex behavior need not have complex causes and that many aspects of irregular, unpredictable phenomena may be understood in terms of simple nonlinear models. However, the study of chaos also teaches us that despite an underlying simplicity and order we will never be able to describe the precise behavior of chaotic systems analytically nor will we succeed in making accurate long-term predictions no matter how much computer power is available. At the very best, we may hope to discern some of this underlying order in an effort to develop reliable statistical methods for making predictions for average properties of chaotic systems.

E. Hyperchaos

The notion of Lyapunov exponent can be extended to systems of differential equations or higher dimensional maps. In general, a system has a set of Lyapunov exponents, each characterizing the average stretching or shrinking of phase space in a particular direction. The logistic map discussed above has only one Lyapunov exponent because it has only one dependent variable that can be displaced. In the Lorenz model, the system has only one *positive* Lyapunov exponent, but has three altogether. Consider an initial point in phase space that is on the strange attractor and

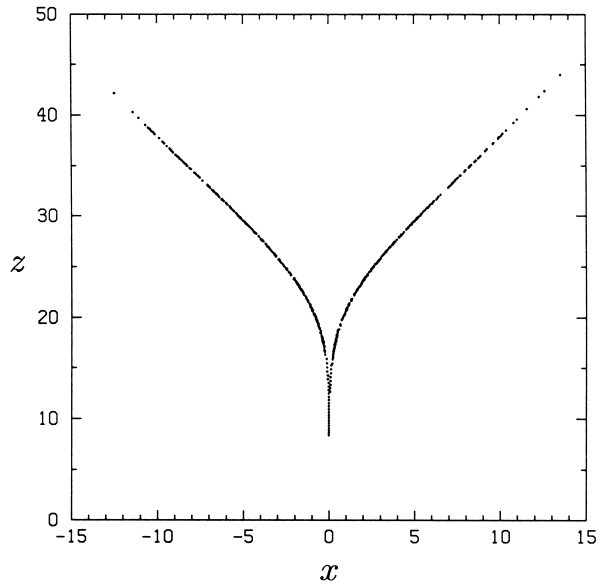


FIGURE 12 A Poincaré section of the Lorenz attractor is constructed by plotting a cross section of the butterfly wings. This graph is generated by plotting the point in the $x-z$ plane each time the orbit displayed in Fig. 11 passes through $y=0$. This view of the strange attractor is analogous to that displayed in Fig. 9 for the Hénon map. This figure appears to consist of only a single strand, but this is because of the large contraction rate of the Lorenz model. Successive magnifications would reveal a fine-scale structure similar to that shown in Fig. 10 for the Hénon map.

another point displaced infinitesimally from it. The trajectories followed from these two points may remain the same distance apart (on average), diverge exponentially, or converge exponentially. The first case corresponds to a Lyapunov exponent of zero and is realized when the displacement lies along the trajectory of the initial point. The two points then follow the same trajectory, but displaced in time. The second case corresponds to a positive Lyapunov exponent, the third to a negative one. In the Lorenz system, the fact that the attractor has a planar structure locally indicates that trajectories converge in the direction transverse to the plane, hence one of the Lyapunov exponents is negative.

An arbitrary, infinitesimal perturbation will almost certainly have some projection on each of the directions corresponding to the different Lyapunov exponents. Since the growth of the perturbation in the direction associated with the largest Lyapunov exponent is exponentially faster than that in any other direction, the observed trajectory divergence will occur in that direction. In numerical models, one can measure the n largest Lyapunov exponents by integrating the linearized equations for the deviations from a given trajectory for n different initial conditions. One must repeatedly rescale the deviations to avoid both exponential growth that causes overflow errors and problems associated with the convergence of all the deviations to the direction associated with the largest exponent.

It is possible to have an attractor with two or more Lyapunov exponents greater than zero. This is sometimes referred to as "hyperchaos" and is common in systems with many degrees of freedom. The problem of distinguishing between hyperchaos and stochastic fluctuations in interpreting experimental data has received substantial attention. We are typically presented with an experimental trace of the time variation of a single variable and wish to determine whether the system that generated it was essentially deterministic or stochastic. The distinction here is quantitative rather than qualitative. If the observed fluctuations involve so many degrees of freedom that it appears hopeless to model them with a simple set of deterministic equations, we label it stochastic and introduce noise terms into the equations.

"Time-series analysis" algorithms have been developed to identify underlying deterministic dynamics in apparently random systems. The central idea behind these algorithms is the construction of a representation of the strange attractor (if it exists) via delay coordinates. Given a time series for a single variable $x(t)$, the n-dimensional vector $\mathbf{X}(t)=(x(t), x(t-\tau), x(t-2\tau), \ldots, x(t-(n-1)\tau))$ is formed, where τ is a fixed delay time comparable to the scale on which $x(t)$ fluctuates. For sufficiently large n, the topological structure of the attractor for $X(t)$ will generically be identical to that of the dynamical system that generated the data. This allows for mesasures of the geometric structure of the trajectory in the space of delay coordinates, called the *embedding space*, to provide an upper bound on the dimension of the true attractor. As a practical matter, hyperchaos with more than about 10 positive Lyapunov exponents is extremely difficult to identify unambiguously.

F. Spatiotemporal Chaos

The term *spatiotemporal chaos* has been used to refer to any system in which some variable exhibits chaotic motion in time and the spatial structure of the system varies with time as well. Real systems are always composed of spatially extended materials for which a fully detailed mathematical model would require either the use of partial differential equations or an enormous number of ordinary differential equations. In many cases, the dynamics of the vast majority of degrees of freedom represented in these equations need not be solved explicitly. All but a few dependent variables exhibit trivial behavior, decaying exponentially quickly to a steady state, or else oscillate at amplitudes negligible for the problem at hand. The remaining variables are described by a few ordinary

differential equations (or maps) of the type discussed in the previous section.

In some cases, the relevant variables are easily identified. For example, it is not difficult to guess that the motion of a pendulum can be described by solving coupled equations for the position and velocity of the bob. One generally does not have to worry about the elastic deformations of the bar. In other cases, the relevant variable are amplitudes of modes of oscillation that can have a nontrivial spatial structure. A system of ordinary differential equations for the amplitudes of a few modes may correspond to an complicated pattern of activity in real space. For example, a violin string can vibrate in different harmonic modes, each of which corresponds to a particular shape of the string that oscillates sinusoidally in time. A model of large-amplitude vibrations might be cast in the form of coupled, nonlinear equations for the amplitudes of a few of the lowest frequency modes. If those equations yielded chaos, the spatial shape of the string would fluctuate in complicated, unpredictable ways. This complex motion in space and time is sometimes referred to as *spatiotemporal chaos*, though it is a rather simple version since the dynamics simplifies greatly when the correct modes are identified.

In general, models of smaller systems require fewer variables in the following sense. What determines the number of modes necessary for an accurate description is the smallest scale spatial variation that has an appreciable probability of occuring at a noticeable amplitude. Since large amplitude variations over very short-length scales generally require large amounts of energy, there will be an effective small-scale cutoff determined by the strength with which the system is driven. Systems whose size is comparable to the cutoff scale will require the analysis of only a few modes; in systems much larger than this scale many modes may be involved and the dynamics can be considerably more complex. The term "spatiotemporal chaos" is sometimes reserved for this regime.

Interest in the general subject of turbulence and its statistical description has led to a number of studies of deterministic systems that exhibit spatiotemporal chaos with a level of complexity proportional to the volume of the system. By analogy with thermodynamic properties that are proportional to the volume, such systems are said to exhibit "extensive chaos." A well-studied example is the irregular pattern of activity known as Benard convection, where a fluid confined to a thin, horizontal layer is heated from below. As the temperature difference between the bottom and top surface of the fluid is increased, the fluid begins to move, arranging itself in a pattern of roughly cylindrical rolls in which warm fluid rises on one side and falls on the other. At the onset of convection, the rolls form straight stripes (apart from boundary effects). As the temperature difference is increased further, the rolls may form more complicated patterns of spirals and defects that continually move around, never settling into a periodic pattern or steady state. The question of whether the "spiral defect chaos" state is an example of extensive chaos is not easy to answer directly, but numerical simulations of models exhibiting similar behavior can be analyzed in detail.

To establish the fact that a numerical model exhibits extensive chaos, one must define an appropriate quantity that characterizes the complexity of the chaotic attractor. A quantity that has proven useful is the Lyapunov dimension, D_λ. Let λ_1 be the largest Lyapunov exponent, λ_2 the second largest, etc. Note that in most extended systems the exponents with higher indices become increasingly strongly negative. Let N be the largest integer for which $\sum_{i=1}^{N} \lambda_i > 0$. We define the Lyapunov dimension as

$$D_\lambda = N + \frac{1}{|\lambda_{N+1}|} \sum_{i=1}^{N} \lambda_i .$$

Numerical studies of systems of partial differential equations such as the complex Ginzburg-Landau equation in two dimensions have demonstrated the existence of attractors for which D_λ does indeed grow proportionally to the system volume; that is, extensive chaos does exist in simple, spatially extended systems, and the spiral defect chaos state is a real example of this phenomenon.

G. Control and Synchronization of Chaos

Over the past 10 years, mathematicians, physicists, and engineers have become increasingly interested in the possibilities of using the unique properties of chaotic systems for novel applications. A key feature of strange attractors that spurred much of this effort was that they have embedded within them an infinite set of perfectly periodic trajectories. These trajectories, called *unstable periodic orbits* (UPOs) lie on the attractor but are not normally observed because they are unstable. In 1990, Ott, Grebogi, and Yorke pointed out that UPOs could form the basis of a switching system. Using standard techniques of control theory for feedback stabilize, we can arrange for an intrinsically chaotic system to follow a selected UPO. By turning off that feedback and turning on a different one, we can stabilize a different UPO. The beauty of the scheme is that we are guaranteed, due to the nature of the strange attractor, that the system will come very close to the desired UPO in a relatively short time. Thus, our feedback system need only be capable of applying tiny perturbations to the system. The chaotic dynamics does the hard work of switching from the vicinity of one orbit to the vicinity of the other.

The notion that chaos can be suppressed using small feedback perturbations has generated a great deal of

interest even independent of the possibility of switching between UPOs. At the time of this writing, applications of "controlling chaos" (or simply suppressing it) are being actively pursued in systems as diverse as semiconductor lasers, mechanical systems, fluid flows, the electrodynamics of cardiac tissue.

A development closely related to controlling chaos has been the use of simple coupling between two nearly identical chaotic systems to synchronize their chaotic behaviors. Given two identical chaotic systems that are uncoupled, their behaviors deviate wildly from each other because of the exponetial divergence of nearby initial conditions. It is possible, however, to couple the systems together in a simple way such that the orginal strange attractor is not altered, but the two systems follow the same trajectory on that attractor. The coupling must be based on the differences between the values of corresponding variables in the two systems. When the systems are synchronized, those differences vanish and the two systems follow the chaotic attractor. If the two systems begin to diverge, however, a feedback is generated via the coupling. An appropriately chosen coupling scheme can maintain the synchronized motion. Synchronization is currently being pursued as a novel means for efficient transmission of information through an electronic or optical channel.

IV. HAMILTONIAN SYSTEMS

Although most physics textbooks on classical mechanics are largely devoted to the description of Hamiltonian systems in which dissipative, frictional forces can be neglected, such systems are rare in nature. The most important examples arise in celestial mechanics, which describes the motions of planets and stars; accelerator design, which deals with tenuous beams of high-energy charged particles moving in guiding magnetic fields; and the physics of magnetically confined plasmas, which is primarily concerned with the dynamics of trapped electrons and ions in high-temperature fusion devices. Although few in number, these examples are very important.

In this section, we will examine three simple examples of classical Hamiltonian systems that exhibit chaotic behavior. The first example is the well-known baker's transformation, which clearly illustrates the fundamental concepts of chaotic behavior in Hamiltonian systems. Although it has no direct applications to physical problems, the baker's transformation, like the logistic map, serves as a paradigm for all chaotic Hamiltonian systems. The second example is the standard map, which has direct applications in the description of the behavior of a wide variety of periodically perturbed nonlinear oscillators ranging from particle motion in accelerators and plasma fusion devices to the irregular rotation of Hyperion, one of the moons of Saturn. Finally, we will consider the Hénon–Heiles model, which corresponds to an autonomous Hamiltonian system with two degrees of freedom, describing, for example, the motion of a particle in a nonaxisymmetric, two-dimensional potential well or the interaction of three nonlinear oscillators (the three-body problem).

A. The Baker's Transformation

The description of a Hamiltonian system, like a frictionless mechanical oscillator, requires at least two dependent variables that usually correspond to a generalized position variable and a generalized momentum variable. These variables define a phase space for the mechanical system, and the solutions of the equations of motion describe the motion of a point in the phase space. Starting from the initial conditions specified by an initial point in the 2–d plane, the time evolution generated by the equations of motion trace out a trajectory or orbit.

The distinctive feature of Hamiltonian systems is that the areas or volumes of small sets of initial conditions are preserved under the time evolution, in contrast to the dissipative systems, such as the Hénon map or the Lorenz model, where phase-space volumes are contracted. Therefore, Hamiltonian systems are not characterized by attractors, either regular or strange, but the dynamics can nevertheless exhibit the same rich variety of behavior with regular periodic and quasi-periodic cycles and chaos.

The simplest Hamiltonian systems correspond to area-preserving maps on the x–y plane. One well-studied example is the so-called baker's transformation, defined by a pair of difference equations:

$$x_{n+1} = 2x_n, \qquad \text{Mod } 1 \tag{15a}$$

$$y_{n+1} = \begin{cases} 0.5y_n & 0 \le x_n \le 0.5 \\ 0.5(y_n + 1) & 0.5 \le x_n \le 1 \end{cases} \tag{15b}$$

The action of this map is easy to describe by using the analogy of how a baker kneads dough (hence, the origin of the name of the map). If we take a set of points (x_n, y_n) covering the unit square ($0 \le x_n \le 1$ and $0 \le y_n \le 1$), Eq. (15a) requires that each value of x_n be doubled so that the square (or dough) is stretched out in the x direction to twice its original length. Then, Eq. (15b) reduces the values of y_n by a factor of two and simultaneously cuts the resulting rectangular set of points (or dough) in half at $x = 1$ and places one piece on top of the other, which returns the dough to its original shape, as shown in Fig. 13. Then, this dynamical process (or kneading) is repeated over and over again.

Since area is preserved under each iteration, this dynamical system is Hamiltonian. This can be easily seen

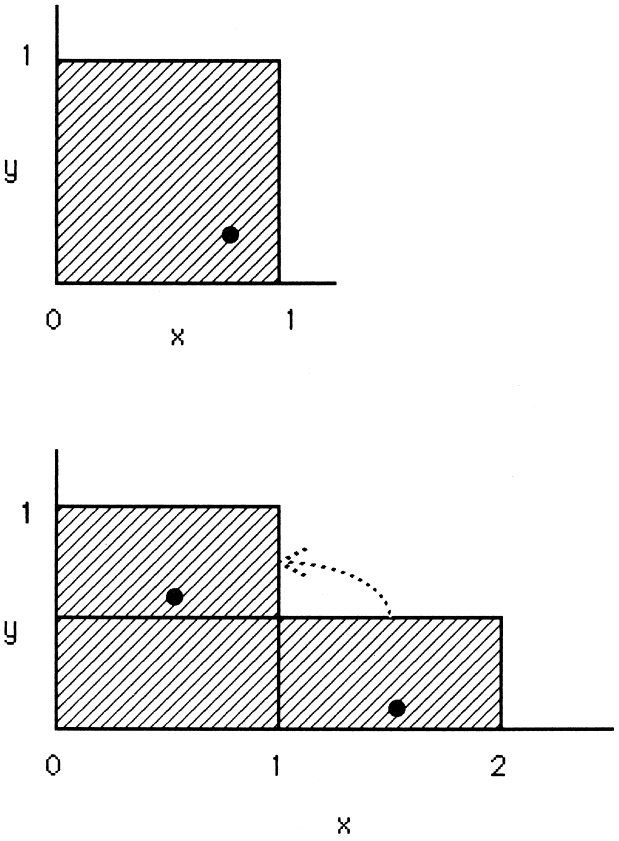


FIGURE 13 The baker's transformation takes all of the points in the unit square (the dough), compresses them vertically by a factor of $\frac{1}{2}$, and stretches them out horizontally by a factor of 2. Then, this rectangular set of points is cut at $x = 1$, the two resulting rectangles are stacked one on top of the other to return the shape to the unit square, and the transformation is repeated over again. In the process, a "raisin," indicated schematically by the black dot, wanders chaotically around the unit square.

mathematically if we think of the successive iterates of the baker's transformation as changes of coordinates from x_n, y_n to x_{n-1}, y_{n+1}. As in the case of the Hénon map, we can analyze the effects of this transformation by evaluating the Jacobian of the coordinate transformation that is the determinant of the matrix:

$$M = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \tag{16}$$

Since $J = \text{Det } M = 1$, we know from elementary integral calculus that volumes are preserved by this change of variables.

1. Chaotic Mixing

Starting from a single initial condition (x_0, y_0), the time evolution will be described by a sequence of points in the plane. (To return to the baking analogy we could imagine that (x_0, y_0) specifies the initial coordinate of a raisin in the dough.) For almost all initial conditions, the trajectories generated by this simple, deterministic map will be chaotic. Because the evolution of the x coordinate is completely determined by the one-dimensional, chaotic shift map, Eq. (2), the trajectory will move from the right half to the left half of the unit square in a sequence that is indistinguishable from the sequence of heads and tails generated by flipping a coin. Moreover, since the location of the orbit in the upper half or lower half of the unit square is determined by the same random sequence, the successive iterates of the initial point (the raisin) will wander around the unit square in a chaotic fashion.

In this simple model, it is easy to see that the mechanism responsible for the chaotic dynamics is the process of stretching and folding of the phase space. In fact, this same stretching and folding lies at the root of all chaotic behavior in both dissipative and Hamiltonian systems. The stretching is responsible for the exponential divergence of nearby trajectories, which is the cause of the extreme sensitivity to initial conditions that characterizes chaotic dynamics. The folding ensures that trajectories return to the initial region of phase space so that the unstable system does not simply explode.

Since the stretching only occurs in the x direction for the baker's transformation, we can easily compute the value of the exponential divergence of nearby trajectories, which is simply the logarithm of the largest eigenvalue of the matrix M. Therefore, the baker's transformation has a positive Kolmogorov–Sinai entropy, $\lambda = \log 2$, so that the dynamics satisfy out definition of chaos.

B. The Standard Map

Our second example of a simple Hamiltonian system that exhibits chaotic behavior is the standard map described by the pair of nonlinear difference equations:

$$x_{n+1} = x_n + y_{n+1}, \qquad \text{Mod } 2\pi \tag{17a}$$

and

$$y_{n+1} = y_n + k \sin x_n \qquad \text{Mod } 2\pi \tag{17b}$$

Starting from an initial point (x_0, y_0) on the "2π square," Eq. (17b) determines the new value of y_1 and Eq. (17a) gives the new value of x_1. The behavior of the trajectory generated by successive iterates is determined by the control parameter k, which measures the strength of the nonlinearity.

The standard map provides a remarkably good model of a wide variety of physical phenomena that are properly described by systems of nonlinear differential equations (hence, the name *standard map*). In particular, it serves as a paradigm for the response of all nonlinear oscillators

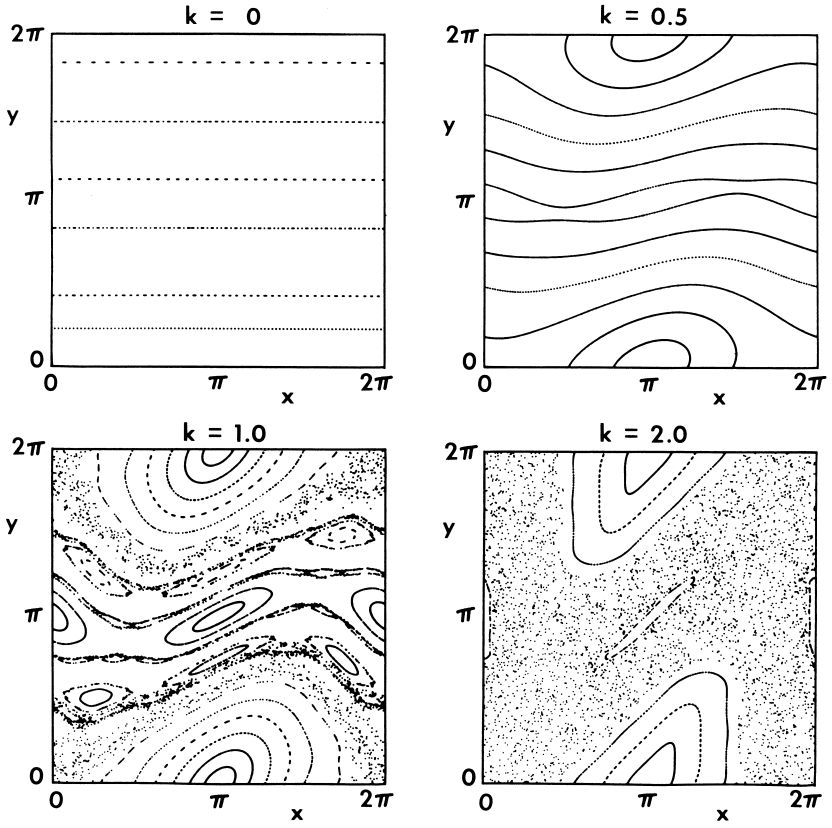


FIGURE 14 Successive iterates of the two-dimensional standard map for a number of different initial conditions are displayed for four values of the control parameter k. For small values of k, the orbits trace out smooth, regular curves in the two-dimensional phase space that become more distorted by resonance effects as k increases. For $k=1$, the interaction of these resonances has generated visible regions of chaos in which individual trajectories wander over large regions of the phase space. However, for $k=1$ and $k=2$, the chaotic regions coexist with regular islets of stability associated with strong nonlinear resonances. The boundaries of the chaotic regions are defined by residual KAM curves. For still larger values of k (not shown), these regular islands shrink until they are no longer visible in these figures.

to periodic perturbations. For example, it provides an approximate description of a particle interacting with a broad spectrum of traveling waves, an electron moving in the imperfect magnetic fields of magnetic bottles used to confine fusion plasmas, and the motion of an electron in a highly excited hydrogen atom in the presence of intense electromagnetic fields. In each case, x_n and y_n correspond to the values of the generalized position and momentum variables, respectively, at discrete times n. Since this model exhibits most of the generic features of Hamiltonian systems that exhibit a transition from regular behavior to chaos, we will examine this example in detail.

The standard map actually provides the exact mathematical description for one physical system called the "kicked rotor." Consider a rigid rotor in the absence of any gravitational or frictional forces that is subject to periodic kicks every unit of time $n = 1, 2, 3, \ldots$. Then, x_n and y_n describe the angle and the angular velocity (angular momentum) just before the nth kick. The rotor can be kicked either forward or backward depending on the sign of $\sin x_n$, and the strengths of the kicks are determined by the value of k. As the nonlinear parameter k is increased, the trajectories generated by this map exhibit a dramatic transition from regular, ordered behavior to chaos. This remarkable transformation is illustrated in Fig. 14, where a number of trajectories are plotted for four different values of k.

When $k = 0$, the value of y remains constant at y_0 and the value of x_n increases each iteration by the amount y_0 (Mod 2π, which means that if x_n does not lie on the interval $[0, 2\pi]$ we add or subtract 2π until it does). In this case, the motion is regular and the trajectories trace out straight lines in the phase space. The rotor rotates continuously at the constant angular velocity y_0. If y_0 is a rational multiple of 2π, then Eq. (17a), like Eq. (1), exhibits a periodic cycle. However, if y_0 is an irrational

multiple of 2π, then the dynamics is quasi-periodic for almost all initial values of x_0 and the points describing the orbit gradually trace out a solid horizontal line in the phase space.

1. Resonance Islands

As k is increased to $k = 0.5$, most of the orbits remain regular and lie on smooth curves in the phase space; however, elliptical islands begin to appear around the point $(\pi, 0) = (\pi, 2\pi)$. (Remember, the intrinsic periodicity of the map implies that the top of the 2π square is connected to the bottom and the right-hand side to the left.) These islands correspond to a resonance between the weak periodic kicks and the rotational frequency of the rotor. Consequently, when the kicks and the rotations are synchronous, the rotor is accelerated. However, because it is a nonlinear oscillator (as opposed to a linear, harmonic oscillator), the rotation frequency changes as the velocity increases so that the motion goes out of resonance and therefore the kicks retard the motion and the velocity decreases until the rotation velocity returns to resonance, then this pattern is repeated. The orbits associated with these quasi-periodic cycles of increasing and decreasing angular velocity trace out elliptical paths in the phase space, as shown in Fig. 14.

The center of the island, $(\pi, 0)$, corresponds to a period-1 point of the standard map. (This is easy to check by simply plugging $(\pi, 0)$ into the right-hand side of Eq. (17).) Figure 14 also shows indications of a smaller island centered at (π, π). Again, it is easy to verify that this point is a member of a period-2 cycle (the other element is the point $(2\pi, \pi) = (0, \pi)$). In fact, there are resonance islands described by chains of ellipses throughout the phase space associated with periodic orbits of all orders. However, most of these islands are much too small to show up in the graphs displayed in Fig. 14.

As the strength of the kicks increases, these islands increase in size and become more prominent. For $k = 1$, several different resonance island chains are clearly visible corresponding to the period-1, period-2, period-3, and period-7 cycles. However, as the resonance regions increase in size and they begin to overlap, individual trajectories between the resonance regions become confused, and the motion becomes chaotic. These chaotic orbits no longer lie on smooth curves in the phase space but begin to wander about larger and larger areas of the phase space as k is increased. For $k = 2$, a single orbit wanders over more than half of the phase space, and for $k = 5$ (not shown), a single orbit would appear to uniformly cover the entire 2π square (although a microscopic examination would always reveal small regular regions near some periodic points).

2. The Kolmogorov–Arnold–Moser Theorem

Since the periodic orbits and the associated resonance regions are mathematically dense in the phase space (though a set of measure zero), there are always small regions of chaos for any nonzero value of k. However, for small values of k, an important mathematical theorem, called the Kolmogorov–Arnold–Moser (KAM) theorem, guarantees that if the perturbation applied to the integrable Hamiltonian system is sufficiently small, then most of the trajectories will lie on smooth curves, such as those displayed in Fig. 14 for $k = 0$ and $k = 0.5$. However, Fig. 14 clearly shows that some of these so-called KAM curves (also called *KAM surfaces* or *invariant tori* in higher dimensions) persist for relatively large values of $k \sim 1$.

The significance of these KAM surfaces is that they form barriers in the phase space. Although these barriers can be circumvented by the slow process of Arnold diffusion in four or more dimensions, they are strictly confining in the two-dimensional phase space of the standard map. This means that orbits starting on one side cannot cross to the other side, and the chaotic regions will be confined by these curves. However, as resonance regions grow with increasing k and begin to overlap, these KAM curves are destroyed and the chaos spreads, as shown in Fig. 14.

The critical k_c for this onset of global chaos can be estimated analytically using Chirikov's resonance overlap criteria, which yields an approximate value of $k_c \approx 2$. However, a more precise value of k_c can be determined by a detailed examination of the breakup of the last confining KAM curve. Since the resonance regions associated with low-order periodic orbits are the largest, the last KAM curve to survive is the one furthest from a periodic orbit. This corresponds to an orbit with the most irrational value of average rotation frequency, which is the golden mean $= (\sqrt{5} - 1)/2$. Careful numerical studies of the standard map show that the golden mean KAM curve, which is the last smooth curve to divide the top of the phase space from the bottom, is destroyed for $k \simeq 1$ (more precisely for $k_c = 0.971635406$). For $k > k_c$, MacKay *et al.* (1987) have shown that this last confining curve breaks up into a so-called cantorus, which is a curve filled with gaps resembling a Cantor set. These gaps allow chaotic trajectories to leak through so that single orbits can wander throughout large regions of the phase space, as shown in Fig. 14 for $k = 2$.

3. Chaotic Diffusion

Because of the intrinsic nonlinearity of Eq. (17b), the restriction of the map to the 2π square was only a graphical convenience that exploited the natural periodicities of the map. However, in reality, both the angle variable and the

angular velocity of a real physical system described by the standard map can take on all real values. In particular, when the golden mean KAM torus is destroyed, the angular velocity associated with the chaotic orbits can wander to arbitrarily large positive and negative values.

Because the chaotic evolution of both the angle and angular velocity appears to execute a random walk in the phase space, it is natural to attempt to describe the dynamics using a statistical description despite the fact that the underlying dynamical equations are fully deterministic. In fact, when $k \gg k_c$, careful numerical studies show that the evolution of an ensemble of initial conditions can be well described by a diffusion equation. Consequently, this simple deterministic dynamical system provides an interesting model for studying the problem of the microscopic foundations of statistical mechanics, which is concerned with the question of how the reversible and deterministic equations of classical mechanics can give rise to the irreversible and statistical equations of classical statistical mechanics and thermodynamics.

C. The Hénon–Heiles Model

Our third example of a Hamiltonian system that exhibits a transition from regular behavior to chaos is described by a system of four coupled, nonlinear differential equations. It was originally introduced by Michel Hénon and Carl Heiles in 1964 as a model of the motion of a star in a nonaxisymmetric, two-dimensional potential corresponding to the mean gravitational field in a galaxy. The equations of motion for the two components of the position and momentum,

$$dx/dt = p_x \tag{18a}$$

$$dy/dt = p_y \tag{18b}$$

$$dp_x/dt = -x - 2xy \tag{18a}$$

$$dp_y/dt = -y + y^2 - x^2 \tag{18b}$$

are generated by the Hamiltonian

$$H(x, y, p_x, p_y) = \frac{p_x^2}{2} + \frac{p_y^2}{2} + \frac{1}{2}(x^2 + y^2) + x^2 y - \frac{1}{3}y^3 \tag{19}$$

where the mass is taken to be unity. Equation 19 corresponds to the Hamiltonian of two uncoupled harmonic oscillators $H_0 = (p_x^2/2) + (p_y^2/2) + \frac{1}{2}(x^2+y^2)$ (consisting of the sum of the kinetic and a quadratic potential energy) plus a cubic perturbation $H_1 = x^2 y - \frac{1}{3}y^3$, which provides a nonlinear coupling for the two linear oscillators.

Since the Hamiltonian is independent of time, it is a constant of motion that corresponds to the total energy of the system $E = H(x, y, p_x, p_y)$. When E is small, both the values of the momenta (p_x, p_y) and the positions (x, y) must remain small. Therefore, in the limit $E \ll 1$, the cubic perturbation can be neglected and the motion will be approximately described by the equations of motion for the unperturbed Hamiltonian, which are easily integrated analytically. Moreover, the application of the KAM theorem to this problem guarantees that as long as E is sufficiently small the motion will remain regular. However, as E is increased, the solutions of the equations of motion, like the orbits generated by the standard map, will become increasingly complicated. First, nonlinear resonances will appear from the coupling of the motions in the x and the y directions. As the energy increases, the effect of the nonlinear coupling grows, the sizes of the resonances grow, and, when they begin to overlap, the orbits begin to exhibit chaotic motion.

1. Poincaré Sections

Although Eq. (18) can be easily integrated numerically for any value of E, it is difficult to graphically display the transition from regular behavior to chaos because the resulting trajectories move in a four-dimensional phase space spanned by x, y, p_x, and p_y. Although we can use the constancy of the energy to reduce the dimension of the accessible phase space to three, the graphs of the resulting three-dimensional trajectories would be even less revealing than the three-dimensional graphs of the Lorenz attractor since there is no attractor to consolidate the dynamics. However, we can simplify the display of the trajectories by exploiting the same device used to relate the Hénon map to the Lorenz model. If we plot the value of p_x versus x every time the orbit passes through $y = 0$, then we can construct a Poincaré section of the trajectory that provides a very clear display of the transition from regular behavior to chaos.

Figure 15 displays these Poincare sections for a number of different initial conditions corresponding to three different energies, $E = \frac{1}{12}$, $\frac{1}{8}$, and $\frac{1}{6}$. For very small E, most of the trajectories lie on an ellipsoid in four-dimensional phase space, so the intersection of the orbits with the p_x–x plane traces out simple ellipses centered at $(x, p_x) = (0, 0)$. For $E = \frac{1}{12}$, these ellipses are distorted and island chains associated with the nonlinear resonances between the coupled motions appear; however, most orbits appear to remain on smooth, regular curves. Finally, as E is increased to $\frac{1}{8}$ and $\frac{1}{6}$, the Poincaré sections reveal a transition from ordered motion to chaos, similar to that observed in the standard map.

In particular, when $E = \frac{1}{6}$, a single orbit appears to uniformly cover most of the accessible phase space defined by the surface of constant energy in the full four-dimensional

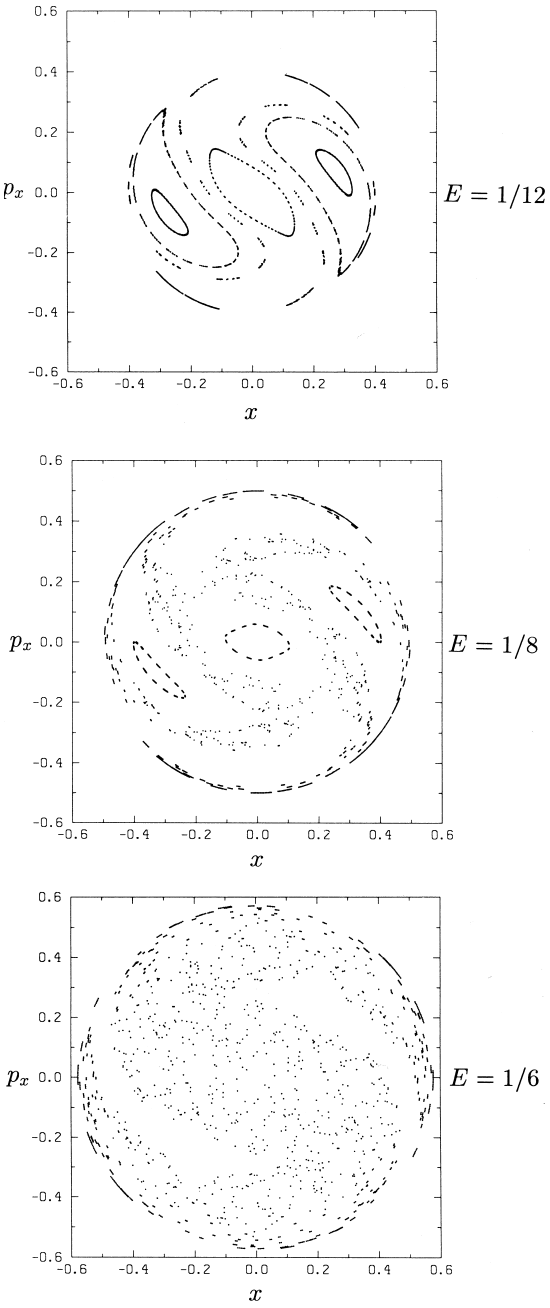


FIGURE 15 Poincaré sections for a number of different orbits generated by the Hénon–Heiles equations are plotted for three different values of the energy E. These figure were created by plotting the position of the orbit in the x–p_x plane each time the solutions of the Hénon–Heiles equations passed through $y = 0$ with positive, p_y. For $E = \frac{1}{12}$, the effect of the perturbation is small and the orbits resemble the smooth but distorted curves observed in the standard map for small k, with resonance islands associated with coupling of the x and y oscillations. However, as the energy increases and the effects of the nonlinearities become more pronounced, large regions of chaotic dynamics become visible and grow until most of the accessible phase space appears to be chaotic for $E = \frac{1}{6}$. (These figures can be compared with the less symmetrical Poincaré sections plotted in the y–p_y plane that usually appear in the literature).

phase space. Although the dynamics of individual trajectories is very complicated in this case, the average properties of an ensemble of trajectories generated by this deterministic but chaotic dynamical system should be well described using the standard methods of statistical mechanics. For example, we may not be able to predict when a star will move chaotically into a particular region of the galaxy, but the average time that the star spends in that region can be computed by simply measuring the relative volume of the corresponding region of the phase space.

D. Applications

The earliest applications of the modern ideas of nonlinear dynamics and chaos to Hamiltonian systems were in the field of accelerator design starting in the late 1950s. In order to maintain a beam of charged particles in an accelerator or storage ring, it is important to understand the dynamics of the corresponding Hamiltonian equations of motion for very long times (in some cases, for more than 10^8 revolutions). For example, the nonlinear resonances associated with the coupling of the radial and vertical oscillations of the beam can be described by models similar to the Hénon–Heiles equations, and the coupling to field oscillations around the accelerator can be approximated by models related to the standard map. In both cases, if the nonlinear coupling or perturbations are too large, the chaotic orbits can cause the beam to defocus and run into the wall.

Similar problems arise in the description of magnetically confined electrons and ions in plasma fusion devices. The densities of these thermonuclear plasmas are sufficiently low that the individual particle motions are effectively collisionless on the time scales of the experiments, so dissipation can be neglected. Again, the nonlinear equations describing the motion of the plasma particles can exhibit chaotic behavior that allows the particles to escape from the confining fields. For example, electrons circulating along the guiding magnetic field lines in a toroidal confinement device called a TOKAMAK will feel a periodic perturbation because of slight variations in magnetic fields, which can be described by a model similar to the standard map. When this perturbation is sufficiently large, electron orbits can become chaotic, which leads to an anomalous loss of plasma confinement that poses a serious impediment to the successful design of a fusion reactor.

The fact that a high-temperature plasma is effectively collisionless also raises another problem in which chaos actually plays a beneficial role and which goes right to the root of a fundamental problem of the microscopic foundations of statistical mechanics. The problem is how do you heat a collisionless plasma? How do you make an irreversible transfer of energy from an external source,

such as the injection of a high-energy particle beam or high-intensity electromagnetic radiation, to a reversible, Hamiltonian system? The answer is chaos. For example, the application of intense radio-frequency radiation induces a strong periodic perturbation on the natural oscillatory motion of the plasma particles. Then, if the perturbation is strong enough, the particle motion will become chaotic. Although the motion remains deterministic and reversible, the chaotic trajectories associated with the ensemble of particles can wander over a large region of the phase space, in particular to higher and lower velocities. Since the temperature is a measure of the range of possible velocities, this process causes the plasma temperature to increase.

Progress in the understanding of chaotic behavior has also caused a revival of interest in a number of problems related to celestial mechanics. In addition to Hénon and Heiles' work on stellar dynamics described previously, Jack Wisdom at MIT has recently solved several old puzzles relating to the origin of meteorites and the presence of gaps in the asteroid belt by invoking chaos. Each time an asteroid that initially lies in an orbit between Mars and Jupiter passes the massive planet Jupiter, it feels a gravitational tug. This periodic perturbation on small orbiting asteroids results in a strong resonant interaction when the two frequencies are related by low-order rational numbers. As in the standard map and the Hénon–Heiles model, if this resonant interaction is sufficiently strong, the asteroid motion can become chaotic. The ideal Kepler ellipses begin to precess and elongate until their orbits cross the orbit of Earth. Then, we see them as meteors and meteorites, and the depletion of the asteroid belts leaves gaps that correspond to the observations.

The study of chaotic behavior in Hamiltonian systems has also found many recent applications in physical chemistry. Many models similar to the Hénon–Heiles model have been proposed for the description of the interaction of coupled nonlinear oscillators that correspond to atoms in a molecule. The interesting questions here relate to how energy is transferred from one part of the molecule to the other. If the classical dynamics of the interacting atoms is regular, then the transfer of energy is impeded by KAM surfaces, such as those in Figs. 14 and 15. However, if the classical dynamics is fully chaotic, then the molecule may exhibit equipartition of energy as predicted by statistical theories. Even more interesting is the common case where some regions of the phase space are chaotic and some are regular. Since most realistic, classical models of molecules involve more than two degrees of freedom, the unraveling of this complex phase-space structure in six or more dimensions remains a challenging problem.

Finally, most recently there has been considerable interest in the classical Hamiltonian dynamics of electrons in highly excited atoms in the presence of strong magnetic fields and intense electromagnetic radiation. The studies of the regular and chaotic dynamics of these strongly perturbed systems have provided a new understanding of the atomic physics in a realm in which conventional methods of quantum perturbation theory fail. However, these studies of chaos in microscopic systems, like those of molecules, have also raised profound, new questions relating to whether the effects of classical chaos can survive in the quantum world. These issues will be discussed in Section V.

V. QUANTUM CHAOS

The discovery that simple nonlinear models of classical dynamical systems can exhibit behavior that is indistinguishable from a random process has naturally raised the question of whether this behavior persists in the quantum realm where the classical nonlinear equations of motion are replaced by the linear Schrodinger equation. This is currently a lively area of research. Although there is general consensus on the key problems, the solutions remain a subject of controversy. In contrast to the subject of classical chaos, there is not even agreement on the definition of quantum chaos. There is only a list of possible symptoms for this poorly characterized disease. In this section, we will briefly discuss the problem of quantum chaos and describe some of the characteristic features of quantum systems that correspond to classically chaotic Hamiltonian systems. Some of these features will be illustrated using a simple model that corresponds to the quantized description of the kicked rotor described in Section IV.B. Then, we will conclude with a description of the comparison of classical and quantum theory with real experiments on highly excited atoms in strong fields.

A. The Problem of Quantum Chaos

Guided by Bohr's correspondence principle, it might be natural to conclude that quantum mechanics should agree with the predictions of classical chaos for macroscopic systems. In addition, because chaos has played a fundamental role in improving our understanding of the microscopic foundations of classical statistical mechanics, one would hope that it would play a similar role in shoring up the foundations of quantum statistical mechanics. Unfortunately, quantum mechanics appears to be incapable of exhibiting the strong local instability that defines classical chaos as a mixing system with positive Kolmogorof–Sinai entropy.

One way of seeing this difficulty is to note that the Schrodinger equation is a linear equation for the wave

function, and neither the wave function nor any observable quantities (determined by taking expectation values of self-adjoint operators) can exhibit extreme sensitivity to initial conditions. In fact, if the Hamiltonian system is bounded (like the Hénon–Heiles Model), then the quantum mechanical energy spectrum is discrete and the time evolution of all quantum mechanical quantities is doomed to quasiperiodic behavior, such as that Eq. (1).

Although the question of the existence of quantum chaos remains a controversial topic, nearly everyone agrees that the most important questions relate to how quantum systems behave when the corresponding classical Hamiltonian systems exhibit chaotic behavior. For example, how does the wave function behave for strongly perturbed oscillators, such as those modeled by the classical standard map, and what are the characteristics of the energy levels for a system of strongly coupled oscillators, such as those described by the Hénon–Heiles model?

B. Symptoms of Quantum Chaos

Even though the Schrödinger equation is a linear equation, the essential nonintegrability of chaotic Hamiltonian systems carries over to the quantum domain. There are no known examples of chaotic classical systems for which the corresponding wave equations can be solved analytically. Consequently, theoretical searches for quantum chaos have also relied heavily on numerical solutions. These detailed numerical studies by physical chemists and physicists studying the dynamics of molecules and the excitation and ionization of atoms in strong fields have led to the identification of several characteristic features of the quantum wave functions and energy levels that reveal the manifestation of chaos in the corresponding classical systems.

One of the most studied characteristics of nonintegrable quantum systems that correspond to classically chaotic Hamiltonian systems is the appearance of irregular energy spectra. The energy levels in the hydrogen atom, described classically by regular, elliptical Kepler orbits, form an orderly sequence, $E_n = -1/(2n^2)$, where $n = 1, 2, 3, \ldots$ is the principal quantum number. However, the energy levels of chaotic systems, such as the quantum Hénon–Heiles model, do not appear to have any simple order at large energies that can be expressed in terms of well-defined quantum numbers. This correspondence makes sense since the quantum numbers that define the energy levels of integrable systems are associated with the classical constants of motion (such as angular momentum), which are destroyed by the nonintegrable perturbation. For example, Fig. 16 displays the calculated energy levels for a hydrogen atom in a magnetic field that shows the transition from the regular spectrum at low magnetic fields to an irregular spectrum ("spaghetti") at high fields in which the magnetic forces are comparable to the Coulomb binding fields.

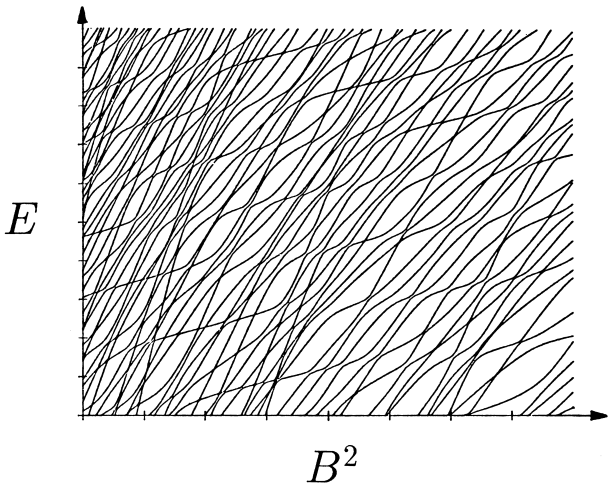


FIGURE 16 The quantum mechanical energy levels for a highly excited hydrogen atom in a strong magnetic field are highly irregular. This figure shows the numerically calculated energy levels as a function of the square of the magnetic field for a range of energies corresponding to quantum states with principal quantum numbers $n \approx 40$–50. Because the magnetic field breaks the natural spherical and Coulomb symmetries of the hydrogen atom, the energy levels and associated quantum states exhibit a jumble of multiple avoided crossings caused by level repulsion, which is a common symptom of quantum systems that are classically chaotic. [From Delande, D. (1988). Ph. D. thesis, Université Pierre & Marie Curie, Paris.]

This irregular spacing of the quantum energy levels can be conveniently characterized in terms of the statistics of the energy level spacings. For example, Fig. 17 shows a histogram of the energy level spacings, $s = E_{i+1} \sim E_i$, for the hydrogen atom in a magnetic field that is strong enough to make most of the classical electron orbits chaotic. Remarkably, this distribution of energy level spacings, $P(s)$, is identical to that found for a much more complicated quantum system with irregular spectra–compound nuclei. Moreover, both distributions are well described by the predictions of random matrix theory, which simply replaces the nonintegrable (or unknown) quantum Hamiltonian with an ensemble of large matrices with random values for the matrix elements. In particular, this distribution of energy level spacings is expected to be given by the Wigner–Dyson distribution, $P(s) \sim s \exp(-s^2)$, displayed in Fig. 17. Although these random matrices cannot predict the location of specific energy levels, they do account for many of the statistical features relating to the fluctuations in the energy level spacings.

Despite the apparent statistical character of the quantum energy levels for classically chaotic systems, these level spacings are not completely random. If they were

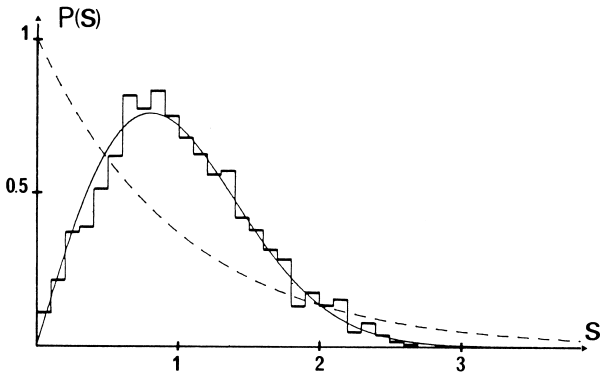


FIGURE 17 The repulsion of the quantum mechanical energy levels displayed in Fig. 16 results in a distribution of energy level spacings, $P(s)$, in which accidental degeneracies ($s=0$) are extremely rare. This figure displays a histogram of the energy level spacings for 1295 levels, such as those in Fig. 16. This distribution compares very well with the Wigner–Dyson distribution (solid curve), which is predicted for the energy level spacing for random matrices. If the energy levels were uncorrelated random numbers, then they would be expected to have a Poisson distribution indicated by the dashed curve. [From Delande. D., and Gay, J. C. (1986). *Phys. Rev. Lett.* **57,** 2006.]

completely uncorrelated, then the spacings statistics would obey a Poison distribution, $P(_s) \sim \exp(s)$, which would predict a much higher probability of nearly degenerate energy levels. The absence of degeneracies in chaotic systems is easily understood because the interaction of all the quantum states induced by the nonintegrable perturbation leads to a repulsion of nearby levels. In addition, the energy levels exhibit an important long-range correlation called *spectral rigidity*, which means that fluctuations about the average level spacing are relatively small over a wide energy range. Michael Berry has traced this spectral rigidity in the spectra of simple chaotic Hamiltonians to the persistence of regular (but not necessarily stable) periodic orbits in the classical phase space. Remarkably, these sets of measure-zero classical orbits appear to have a dominant influence on the characteristics of the quantum energy levels and quantum states.

Experimental studies of the energy levels of Rydberg atoms in strong magnetic fields by Karl Welge and collaborators at the University of Bielefeld appear to have confirmed many of these theoretical and numerical predictions. Unfortunately, the experiments can only resolve a limited range of energy levels, which makes the confirmation of statistical predictions difficult. However, the experimental observations of this symptom of quantum chaos are very suggestive. In addition, the experiments have provided very striking evidence for the important role of classical regular orbits embedded in the chaotic sea of trajectories in determining gross features in the fluctuations in the irregular spectrum. In particular, there appears to be a one-to-one correspondence between regular oscillations in the spectrum and the periods of the shortest periodic orbits in the classical Hamiltonian system. Although the corresponding classical dynamics of these simple systems is fully chaotic, the quantum mechanics appears to cling to the remnants of regularity.

Another symptom of quantum chaos that is more direct is to simply look for quantum behavior that resembles the predictions of classical chaos. In the cases of atoms or molecules in strong electromagnetic fields where classical chaos predicts ionization or dissociation, this symptom is unambiguous. (The patient dies.) However, quantum systems appear to be only capable of mimicking classical chaotic behavior for finite times determined by the density of quantum states (or the size of the quantum numbers). In the case of as few as 50 interacting particles, this break time may exceed the age of the universe, however, for small quantum systems, such as those described by the simple models of Hamiltonian chaos, this time scale, where the Bohr correspondence principle for chaotic systems breaks down, may be accessible to experimental measurements.

C. The Quantum Standard Map

One model system that has greatly enhanced our understanding of the quantum behavior of classically chaotic systems is the quantum standard map, which was first introduced by Casati *et al*. in 1979. The Schrodinger equation for the kicked rotor described in Section IV.B also reduces to a map that describes how the wave function (expressed in terms of the unperturbed quantum eigenstates of the rotor) spreads at each kick. Although this map is formally described by an infinite system of linear difference equations, these equations can be solved numerically to good approximation by truncating the set of equations to a large but finite number (typically, $\leq$1000 states).

The comparison of the results of these quantum calculations with the classical results for the evolution of the standard map over a wide range of parameters has revealed a number of striking features. For short times, the quantum evolution resembles the classical dynamics generated by evolving an ensemble of initial conditions with the same initial energy or angular momenta but different initial angles. In particular, when the classical dynamics is chaotic, the quantum mechanical average of the kinetic energy also increases linearly up to a break time where the classical dynamics continue to diffuse in angular velocity but the quantum evolution freezes and eventually exhibits quasi-periodic recurrences to the initial state. Moreover, when the classical mechanics is regular the quantum wave function is also confined by the KAM surfaces for short times but may eventually "tunnel" or leak through.

This relatively simple example shows that quantum mechanics is capable of stabilizing the dynamics of the classically chaotic systems and destabilizing the regular classical dynamics, depending on the system parameters. In addition, this dramatic quantum suppression of classical chaos in the quantum standard map has been related to the phenomenon of Anderson localization in solid-state physics where an electron in a disordered lattice will remain localized (will not conduct electricity) through destructive quantum interference effects. Although there is no random disorder in the quantum standard map, the classical chaos appears to play the same role.

D. Microwave Ionization of Highly Excited Hydrogen Atoms

As a consequence of these suggestive results for the quantum standard map, there has been a considerable effort to see whether the manifestations of classical chaos and its suppression by quantum interference effects could be observed experimentally in a real quantum system consisting of a hydrogen atom prepared in a highly excited state that is then exposed to intense microwave fields.

Since the experiments can be performed with atoms prepared in states with principal quantum numbers as high as $n = 100$, one could hope that the dynamics of this electron with a 0.5-μm Bohr radius would be well described by classical dynamics. In the presence of an intense oscillating field, this classical nonlinear oscillator is expected to exhibit a transition to global chaos such as that exhibited by the classical standard map at $k \approx 1$. For example, Fig. 18 shows a Poincaré section of the classical action-angle phase space for a one-dimensional model of a hydrogen atom in an oscillating field for parameters that correspond closely to those of the experiments. For small values of the classical action I, which correspond to low quantum numbers by the Bohr–Somerfeld quantization rule, the perturbing field is much weaker than the Coulomb binding fields and the orbits lie on smooth curves that are bounded by invariant KAM tori. However, for larger values of I, the relative size of the perturbation increases and the orbits become chaotic, filling large regions of phase space and wandering to arbitrarily large values of the action and ionizing. Since these chaotic orbits ionize, the classical theory predicts an ionization mechanism that depends strongly on the intensity of the radiation and only weakly on the frequency, which is just the opposite of the dependence of the traditional photoelectric effect.

In fact, this chaotic ionization mechanism was first experimentally observed in the pioneering experiments of Jim Bayfield and Peter Koch in 1974, who observed the sharp onset of ionization in atoms prepared in the $n \approx 66$ state, when a 10-GHz microwave field exceeded a critical threshold. Subsequently, the agreement of the predictions

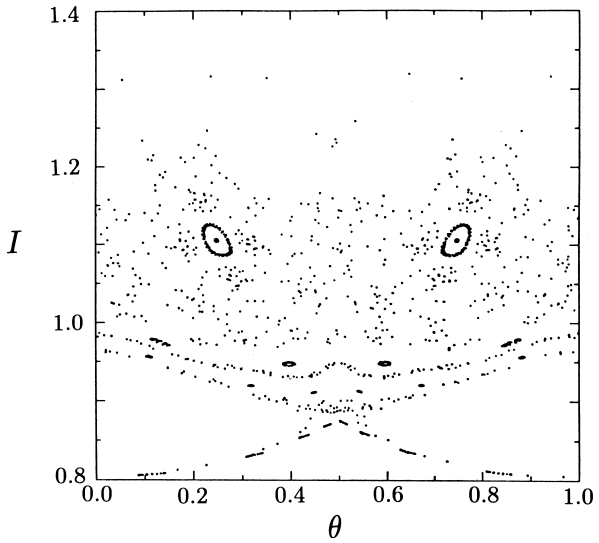


FIGURE 18 This Poincaré section of the classical dynamics of a one-dimensional hydrogen atom in a strong oscillating electric field was generated by plotting the value of the classical action I and angle θ once every period of the perturbation with strength $I^4F = 0.03$ and frequency $I^3\Omega = 1.5$. In the absence of the perturbations, the action (which corresponds to principal quantum number n by the Bohr–Sommerfeld quantization rule) is a constant of motion. In this case, different initial conditions (corresponding to different quantum states of the hydrogen atom) would trace out horizontal lines in the phase space, such as those in Fig. 14, for the standard map at $k = 0$. Since the Coulomb binding field decreases as $1/I^4$ (or $1/n^4$), the relative strength of the perturbation increases with I. For a fixed value of the perturbing field F, the classical dynamics is regular for small values of I with a prominent nonlinear resonance below $I = 1.0$. A prominent pair of islands also appears near $I = 1.1$, but it is surrounded by a chaotic sea. Since the chaotic orbits can wander to arbitrarily high values of the action, they ultimately led to ionization of the atom.

of classical chaos on the quantum measurements has been confirmed for a wide range of parameters corresponding to principal quantum numbers from $n = 32$ to 90. Figure 19 shows the comparison of the measured thresholds for the onset of ionization with the theoretical predictions for the onset of classical chaos in a one-dimensional model of the experiment.

Moreover, detailed numerical studies of the solution of the Schrödinger equation for the one-dimensional model have revealed that the quantum mechanism that mimics the onset of classical chaos is the abrupt delocalization of the evolving wave packet when the perturbation exceeds a critical threshold. However, these quantum calculations also showed that in a parameter range just beyond that studied in the original experiments the threshold fields for quantum delocalization would become larger than the classical predictions for the onset of chaotic ionization. This quantum suppression of the classical chaos would be analogous to that observed in the quantum standard map. Very recently, the experiments in this new regime

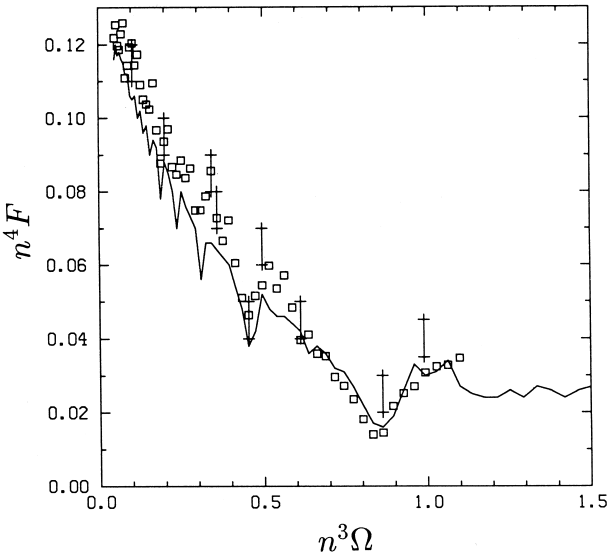


FIGURE 19 A comparison of the threshold field strengths for the onset of microwave ionization predicted by the classical theory for the onset of chaos (solid curve) with the results of experimental measurements on real hydrogen atoms with $n = 32$ to 90 (open squares) and with estimates from the numerical solution of the corresponding Schrödinger equation (crosses). The threshold field strengths are conveniently plotted in terms of the scaled variable $n^4F = I^4F$, which is the ratio of the perturbing field F to the Coulomb binding field $1/n^4$ versus the scaled frequency $n^3\Omega = I^3\Omega$, which is the ratio of the microwave frequency Ω to the Kepler orbital frequency $1/n^3$. The prominent features near rational values of the scaled frequency, $n^3\Omega = 1$, $\frac{1}{2}$, $\frac{1}{3}$, *and* $\frac{1}{4}$, which appear in both the classical and quantum calculations as well as the experimental measurements, are associated with the presence of nonlinear resonances in the classical phase space.

have been performed, and the experimental evidence supports the theoretical prediction for quantum suppression of classical chaos, although the detailed mechanisms remain a topic of controversy.

These experiments and the associated classical and quantum theories are parts of the exploration of the frontiers of a new regime of atomic and molecular physics for strongly interacting and strongly perturbed systems. As our understanding of the dynamics of the simplest quantum systems improves, these studies promise a number of important applications to problems in atomic and molecular physics, physical chemistry, solid-state physics, and nuclear physics.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTIC CHAOS • ATOMIC AND MOLECULAR COLLISIONS • COLLIDER DETECTORS FOR MULTI-TEV PARTICLES • FLUID DYNAMICS • FRACTALS • MATHEMATICAL MODELING • MECHANICS, CLASSICAL • NONLINEAR DYNAMICS • QUANTUM THEORY • TECTONOPHYSICS • VIBRATION, MECHANICAL

BIBLIOGRAPHY

Baker, G. L., and Gollub, J. P. (1990). "Chaotic Dynamics: An Introduction," Cambridge University Press, New York.

Berry, M. V. (1983). "Semi-classical mechanics of regular and irregular motion," In "Chaotic Behavior of Deterministic Systems" (G. Iooss, R. H. G. Helleman, and R. H. G. Stora, eds.), p. 171. North-Holland, Amsterdam.

Berry, M. V. (1985). "Semi-classical theory of spectral rigidity," *Proc. R. Soc. Lond. A* **400,** 229.

Bohr, T., Jensen, M. H., Paladin, G., and Vulpiani, A. (1998). "Dynamical Systems Approach to Turbulence," Cambridge University Press, New York.

Campbell, D., ed. (1983). "Order in Chaos, Physica 7D," Plenum, New York.

Casati, G., ed. (1985). "Chaotic Behavior in Quantum Systems," Plenum, New York.

Casati, G., Chirikov, B. V., Shepelyansky, D. L., and Guarneri, I. (1987). "Relevance of classical chaos in quantum mechanics: the hydrogen atom in a monochromatic field," *Phys. Rep*. **154,** 77.

Crutchfield, J. P., Farmer, J. D., Packard, N. H., and Shaw, R. S. (1986). "Chaos," *Sci. Am*. **255,** 46.

Cvitanovic, P., ed. (1984). "Universality in Chaos," Adam Hilger, Bristol. (This volume contains a collection of the seminal articles by M. Feigenbaum, E. Lorenz, R. M. May, and D. Ruelle, as well as an excellent review by R. H. G. Helleman.)

Ford, J. (1983). "How random is a coin toss?" *Phys. Today* **36,** 40.

Giannoni, M.-J., Voros, A., and Zinn-Justin, J., eds. (1990). "Chaos and Quantum Physics," Elsevier Science, London.

Gleick, J. (1987). "Chaos: Making of a New Science," Viking, New York.

Gutzwiller, M. C. (1990). "Choas in Classical and Quantum Mechanics," Springer-Verlag, New York. (This book treats the correspondence between classical chaos and relevant quantum systems in detail, on a rather formal level.)

Jensen, R. V. (1987a). "Classical chaos," *Am. Sci*. **75,** 166.

Jensen, R. V. (1987b). "Chaos in atomic physics," *In* "Atomic Physics 10" (H. Narami and I. Shimimura, eds.), p. 319, North-Holland, Amsterdam.

Jensen, R. V. (1988). "Chaos in atomic physics," *Phys. Today* **41,** S-30.

Jensen, R. V., Susskind, S. M., and Sanders, M. M. (1991). "Chaotic ionization of highly excited hydrogen atoms: comparison of classical and quantum theory with experiment," *Phys. Rep*. **201,** 1.

Lichtenberg, A. J., and Lieberman, M. A. (1983). "Regular and Stochastic Motion," Springer-Verlag, New York.

MacKay, R. S., and Meiss, J. D., eds. (1987). "Hamiltonian Dynamical Systems," Adam Hilger, Bristol.

Mandelbrot, B. B. (1982). "The Fractal Geometry of Nature," Freeman, San Francisco.

Ott, E. (1981). "Strange attractors and chaotic motions off dynamical systems," *Rev. Mod. Phys.* **53,** 655.

Ott, E. (1993). "Chaos in Dynamical Systems," Cambridge University Press, New York. (This is a comprehensive, self-contained introduction to the subject of chaos, presented at a level appropriate for graduate students and researchers in the physical sciences, mathematics, and engineering.)

Physics Today (1985). "Chaotic orbits and spins in the solar system," *Phys. Today* **38,** 17.

Schuster, H. G. (1984). "Deterministic Chaos," Physik-Verlag, Weinheim, F. R. G.



Charged-Particle Optics

P. W. Hawkes
CNRS, Toulouse, France

I. Introduction
II. Geometric Optics
III. Wave Optics
IV. Concluding Remarks

GLOSSARY

Aberration A perfect lens would produce an image that was a scaled representation of the object; real lenses suffer from defects known as aberrations and measured by aberration coefficients.

Cardinal elements The focusing properties of optical components such as lenses are characterized by a set of quantities known as cardinal elements; the most important are the positions of the foci and of the principal planes and the focal lengths.

Conjugate Planes are said to be conjugate if a sharp image is formed in one plane of an object situated in the other. Corresponding points in such pairs of planes are also called conjugates.

Electron lens A region of space containing a rotationally symmetric electric or magnetic field created by suitably shaped electrodes or coils and magnetic materials is known as a round (electrostatic or magnetic) lens. Other types of lenses have lower symmetry; quadrupole lenses, for example, have planes of symmetry or antisymmetry.

Electron prism A region of space containing a field in which a plane but not a straight optic axis can be defined forms a prism.

Image processing Images can be improved in various ways by manipulation in a digital computer or by optical analog techniques; they may contain latent information, which can similarly be extracted, or they may be so complex that a computer is used to reduce the labor of analyzing them. Image processing is conveniently divided into acquisition and coding; enhancement; restoration; and analysis.

Optic axis In the optical as opposed to the ballistic study of particle motion in electric and magnetic fields, the behavior of particles that remain in the neighborhood of a central trajectory is studied. This central trajectory is known as the optic axis.

Paraxial Remaining in the close vicinity of the optic axis. In the paraxial approximation, all but the lowest order terms in the general equations of motion are neglected, and the distance from the optic axis and the gradient of the trajectories are assumed to be very small.

Scanning electron microscope (SEM) Instrument in which a small probe is scanned in a raster over the surface of a specimen and provokes one or several signals, which are then used to create an image on a cathoderay tube or monitor. These signals may be X-ray intensities or secondary electron or backscattered electron currents, and there are several other possibilities.

Scanning transmission electron microscope (STEM) As in the scanning electron microscope, a small probe explores the specimen, but the specimen is thin and the signals used to generate the images are detected downstream. The resolution is comparable with that of the transmission electron microscope.

Scattering When electrons strike a solid target or pass through a thin object, they are deflected by the local field. They are said to be scattered, elastically if the change of direction is affected with negligible loss of energy, inelastically when the energy loss is appreciable.

Transmission electron microscope (TEM) Instrument closely resembling a light microscope in its general principles. A specimen area is suitably illuminated by means of condenser lenses. An objective close to the specimen provides the first stage of magnification, and intermediate and projector lens magnify the image further. Unlike glass lenses, the lens strength can be varied at will, and the total magnification can hence be varied from a few hundred times to hundreds of thousands of times. Either the object plane or the plane in which the diffraction pattern of the object is formed can be made conjugate to the image plane.

OF THE MANY PROBES used to explore the structure of matter, charged particles are among the most versatile. At high energies they are the only tools available to the nuclear physicist; at lower energies, electrons and ions are used for high-resolution microscopy and many related tasks in the physical and life sciences. The behavior of the associated instruments can often be accurately described in the language of optics. When the wavelength associated with the particles is unimportant, geometric optics are applicable and the geometric optical properties of the principal optical components—round lenses, quadrupoles, and prisms—are therefore discussed in detail. Electron microscopes, however, are operated close to their theoretical limit of resolution, and to understand how the image is formed a knowledge of wave optics is essential. The theory is presented and applied to the two families of high-resolution instruments.

I. INTRODUCTION

Charged particles in motion are deflected by electric and magnetic fields, and their behavior is described either by the Lorentz equation, which is Newton's equation of motion modified to include any relativistic effects, or by Schrödinger's equation when spin is negligible. There are many devices in which charged particles travel in a restricted zone in the neighborhood of a curve, or axis, which is frequently a straight line, and in the vast majority of these devices, the electric or magnetic fields exhibit some very simple symmetry. It is then possible to describe the deviations of the particle motion by the fields in the familiar language of optics. If the fields are rotationally symmetric about an axis, for example, their effects are closely analogous to those of round glass lenses on light rays. Focusing can be described by cardinal elements, and the associated defects resemble the geometric and chromatic aberrations of the lenses used in light microscopes, telescopes, and other optical instruments. If the fields are not rotationally symmetric but possess planes of symmetry or antisymmetry that intersect along the optic axis, they have an analog in toric lenses, for example the glass lenses in spectacles that correct astigmatism. The other important field configuration is the analog of the glass prism; here the axis is no longer straight but a plane curve, typically a circle, and such fields separate particles of different energy or wavelength just as glass prisms redistribute white light into a spectrum.

In these remarks, we have been regarding charged particles as classical particles, obeying Newton's laws. The mention of wavelength reminds us that their behavior is also governed by Schrödinger's equation, and the resulting description of the propagation of particle beams is needed to discuss the resolution of electron-optical instruments, notably electron microscopes, and indeed any physical effect involving charged particles in which the wavelength is not negligible.

Charged-particle optics is still a young subject. The first experiments on electron diffraction were made in the 1920s, shortly after Louis de Broglie associated the notion of wavelength with particles, and in the same decade Hans Busch showed that the effect of a rotationally symmetric magnetic field acting on a beam of electrons traveling close to the symmetry axis could be described in optical terms. The first approximate formula for the focal length was given by Busch in 1926–1927. The fundamental equations and formulas of the subject were derived during the 1930s, with Walter Glaser and Otto Scherzer contributing many original ideas, and by the end of the decade the German Siemens Company had put the first commercial electron microscope with magnetic lenses on the market. The latter was a direct descendant of the prototypes built by Max Knoll, Ernst Ruska, and Bodo von Borries from 1932 onwards. Comparable work on the development of an electrostatic instrument was being done by the AEG Company.

Subsequently, several commercial ventures were launched, and French, British, Dutch, Japanese, Swiss,

American, Czechoslovakian, and Russian electron microscopes appeared on the market as well as the German instruments. These are not the only devices that depend on charged-particle optics, however. Particle accelerators also use electric and magnetic fields to guide the particles being accelerated, but in many cases these fields are not static but dynamic; frequently the current density in the particle beam is very high. Although the traditional optical concepts need not be completely abandoned, they do not provide an adequate representation of all the properties of "heavy" beams, that is, beams in which the current density is so high that interactions between individual particles are important. The use of very high frequencies likewise requires different methods and a new vocabulary that, although known as "dynamic electron optics," is far removed from the optics of lenses and prisms. This account is confined to the charged-particle optics of static fields or fields that vary so slowly that the static equations can be employed with negligible error (scanning devices); it is likewise restricted to beams in which the current density is so low that interactions between individual particles can be neglected, except in a few local regions (the crossover of electron guns).

New devices that exploit charged-particle optics are constantly being added to the family that began with the transmission electron microscope of Knoll and Ruska. Thus, in 1965, the Cambridge Instrument Co. launched the first commercial scanning electron microscope after many years of development under Charles Oatley in the Cambridge University Engineering Department. Here, the image is formed by generating a signal at the specimen by scanning a small electron probe over the latter in a regular pattern and using this signal to modulate the intensity of a cathode-ray tube. Shortly afterward, Albert Crewe of the Argonne National Laboratory and the University of Chicago developed the first scanning transmission electron microscope, which combines all the attractions of a scanning device with the very high resolution of a "conventional" electron microscope. More recently still, fine electron beams have been used for microlithography, for in the quest for microminiaturization of circuits, the wavelength of light set a lower limit on the dimensions attainable. Finally, there are, many devices in which the charged particles are ions of one or many species. Some of these operate on essentially the same principles as their electron counterparts; in others, such as mass spectrometers, the presence of several ion species is intrinsic. The laws that govern the motion of all charged particles are essentially the same, however, and we shall consider mainly electron optics; the equations are applicable to any charged particle, provided that the appropriate mass and charge are inserted.

II. GEOMETRIC OPTICS

A. Paraxial Equations

Although it is, strictly speaking, true that any beam of charged particles that remains in the vicinity of an arbitrary curve in space can be described in optical language, this is far too general a starting point for our present purposes. Even for light, the optics of systems in which the axis is a skew curve in space, developed for the study of the eye by Allvar Gullstrand and pursued by Constantin Carathéodory, are little known and rarely used. The same is true of the corresponding theory for particles, developed by G. A. Grinberg and Peter Sturrock. We shall instead consider the other extreme case, in which the axis is straight and any magnetic and electrostatic fields are rotationally symmetric about this axis.

1. Round Lenses

We introduce a Cartesian coordinate system in which the z axis coincides with the symmetry axis, and we provisionally denote the transverse axes X and Y. The motion of a charged particle of rest mass m_0 and charge Q in an electrostatic field $\mathbf{E}$ and a magnetic field $\mathbf{B}$ is then determined by the differential equation

$$(d/dt)(\gamma m_0 \mathbf{v}) = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$
$$\gamma = (1 - v^2/c^2)^{-1/2}, \tag{1}$$

which represents Newton's second law modified for relativistic effects (Lorentz equation); $\mathbf{v}$ is the velocity. For electrons, we have $e = -Q \simeq 1.6 \times 10^{-19}$ C and $e/m_0 \simeq 176$ C/μg. Since we are concerned with static fields, the time of arrival of the particles is often of no interest, and it is then preferable to differentiate not with respect to time but with respect to the axial coordinate z. A fairly lengthy calculation yields the trajectory equations

$$\frac{d^2X}{dz^2} = \frac{\rho^2}{g}\left(\frac{\partial g}{\partial X} - X'\frac{\partial g}{\partial z}\right) + \frac{Q\rho}{g}\left[Y'(B_z + X'B_X) - B_Y(1 + X'^2)\right]$$
$$\frac{d^2Y}{dz^2} = \frac{\rho^2}{g}\left(\frac{\partial g}{\partial Y} - Y'\frac{\partial g}{\partial z}\right) + \frac{Q\rho}{g}\left[-X'(B_z + Y'B_Y) + B_X(1 + Y'^2)\right] \tag{2}$$

in which $\rho^2 = 1 + X'^2 + Y'^2$ and $g = \gamma m_0 v$.

By specializing these equations to the various cases of interest, we obtain equations from which the optical properties can be derived by the "trajectory method." It is well

known that equations such as Eq. (1) are identical with the Euler–Lagrange equations of a variational principle of the form

$$W = \int_{t_0}^{t_1} L(\mathbf{r}, \mathbf{v}, t)\, dt = \text{extremum} \tag{3}$$

provided that t_0, t_1, $\mathbf{r}(t_0)$, and $\mathbf{r}(t_1)$ are held constant. The Lagrangian L has the form

$$L = m_0 c^2[1 - (1 - v^2/c^2)^{1/2}] + Q(\mathbf{v} \cdot \mathbf{A} - \Phi) \tag{4}$$

in which Φ and $\mathbf{A}$ are the scalar and vector potentials corresponding to $\mathbf{E}$, $\mathbf{E} = -\text{grad}\ \Phi$ and to $\mathbf{B}$, $\mathbf{B} = \text{curl}\ \mathbf{A}$. For static systems with a straight axis, we can rewrite Eq. (3) in the form

$$S = \int_{z_0}^{z_1} M(x, y, z, x', y')\, dz, \tag{5}$$

where

$$M = (1 + X'^2 + Y'^2)^{1/2} g(\mathbf{r}) + Q(X'A_X + Y'A_Y + A_z). \tag{6}$$

The Euler–Lagrange equations,

$$\frac{d}{dz}\left(\frac{\partial M}{\partial X'}\right) = \frac{\partial M}{\partial X}; \qquad \frac{d}{dz}\left(\frac{\partial M}{\partial Y'}\right) = \frac{\partial M}{\partial Y} \tag{7}$$

again define trajectory equations. A very powerful method of analyzing optical properties is based on a study of the function M and its integral S; this is known as the method of characteristic functions, or eikonal method.

We now consider the special case of rotationally symmetric systems in the paraxial approximation; that is, we examine the behavior of charged particles, specifically electrons, that remain very close to the axis. For such particles, the trajectory equations collapse to a simpler form, namely,

$$\begin{aligned} X'' + \frac{\gamma\phi'}{2\hat{\phi}}X' + \frac{\gamma\phi''}{4\hat{\phi}}X + \frac{\eta B}{\hat{\phi}^{1/2}}Y' + \frac{\eta B'}{2\hat{\phi}^{1/2}}Y = 0 \\ Y'' + \frac{\gamma\phi'}{2\hat{\phi}}Y' + \frac{\gamma\phi''}{4\hat{\phi}}Y - \frac{\eta B}{\hat{\phi}^{1/2}}X' - \frac{\eta B'}{2\hat{\phi}^{1/2}}X = 0 \end{aligned} \tag{8}$$

in which $\phi(z)$ denotes the distribution of electrostatic potential on the optic axis, $\phi(z) = \Phi(0, 0, z)$; $\hat{\phi}(z) = \phi(z)[1 + e\phi(z)/2m_0c^2]$. Likewise, $B(z)$ denotes the magnetic field distribution on the axis. These equations are coupled, in the sense that X and Y occur in both, but this can be remedied by introducing new coordinate axes x, y, inclined to X and Y at an angle $\theta(z)$ that varies with z; $x = 0$, $y = 0$ will therefore define not planes but surfaces. By choosing $\theta(z)$ such that

$$d\theta/dz = \eta B/2\hat{\phi}^{1/2}; \qquad \eta = (e/2m_0)^{1/2}, \tag{9}$$

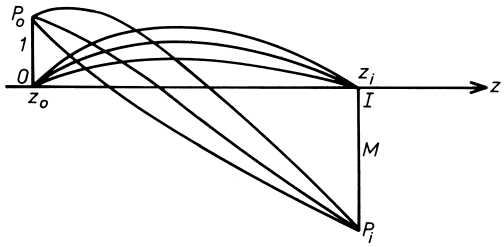


FIGURE 1 Paraxial solutions demonstrating image formation.

we find

$$\begin{aligned} x'' + \gamma\phi' x'/2\hat{\phi} + [(\gamma\phi'' + \eta^2 B^2)/4\hat{\phi}]/x = 0 \\ y'' + \gamma\phi' y'/2\hat{\phi} + [(\gamma\phi'' + \eta^2 B^2)/4\hat{\phi}]/y = 0. \end{aligned} \tag{10}$$

These differential equations are linear, homogeneous, and second order. The general solution of either is a linear combination of any two linearly independent solutions, and this fact is alone sufficient to show that the corresponding fields $B(z)$ and potentials $\phi(z)$ have an imaging action, as we now show. Consider the particular solution $h(z)$ of Eq. (10) that intersects the axis at $z = z_0$ and $z = z_i$ (Fig. 1). A pencil of rays that intersects the plane $z = z_o$ at some point $P_o(x_o, y_o)$ can be described by

$$\begin{aligned} x(z) = x_o g(z) + \lambda h(z) \\ y(z) = y_o g(z) + \mu h(z) \end{aligned} \tag{11}$$

in which $g(z)$ is any solution of Eq. (10) that is linearly independent of $h(z)$ such that $g(z_o) = 1$ and λ, μ are parameters; each member of the pencil corresponds to a different pair of values of λ, μ. In the plane $z = z_i$, we find

$$x(z_i) = x_o g(z_i); \qquad y(z_i) = y_o g(z_i) \tag{12}$$

for all λ and μ and hence for all rays passing through P_o. This is true for every point in the plane $z = z_o$, and hence the latter will be stigmatically imaged in $z = z_i$.

Furthermore, both ratios and $x(z_i)/x_o$ and $y(z_i)/y_o$ are equal to the constant $g(z_i)$, which means that any pattern of points in $z = z_o$ will be reproduced faithfully in the image plane, magnified by this factor $g(z_i)$, which is hence known as the (transverse) magnification and denoted by M.

The form of the paraxial equations has numerous other consequences. We have seen that the coordinate frame x–y–z rotates relative to the fixed frame X–Y–Z about the optic axis, with the result that the image will be rotated with respect to the object if magnetic fields are used. In an instrument such as an electron microscope, the image therefore rotates as the magnification is altered, since the latter is affected by altering the strength of the magnetic field and Eq. (9) shows that the angle of rotation is a function of this quantity. Even more important is the fact that the coefficient of the linear term is strictly positive in the

case of magnetic fields. This implies that the curvature of any solution $x(z)$ is opposite in sign to $x(z)$, with the result that the field always drives the electrons toward the axis; magnetic electron lenses always have a convergent action. The same is true of the overall effect of electrostatic lenses, although the reasoning is not quite so simple.

A particular combination of any two linearly independent solutions of Eq. (10) forms the invariant known as the Wronskian. This quantity is defined by

$$\hat{\phi}^{1/2}(x_1x_2' - x_1'x_2); \qquad \hat{\phi}^{1/2}(y_1y_2' - y_1'y_2) \tag{13}$$

Suppose that we select $x_1 = h$ and $x_2 = g$, where $h(z_o) = h(z_i) = 0$ and $g(z_o) = 1$ so that $g(z_i) = M$. Then

$$\hat{\phi}_o^{1/2}h_o' = \hat{\phi}_i^{1/2}h_i'M \tag{14}$$

The ratio h_i'/h_o' is the angular magnification M_A and so

$$MM_A = (\hat{\phi}_o/\hat{\phi}_i)^{1/2} \tag{15}$$

or $MM_A = 1$ if the lens has no overall accelerating effect and hence $\hat{\phi}_o = \hat{\phi}_i$. Identifying $\phi^{1/2}$ with the refractive index, Eq. (15) is the particle analog of the Smith–Helmholtz formula of light optics. Analogs of all the other optical laws can be established; in particular, we find that the longitudinal magnification M_l is given by.

$$M_l = M/M_A = (\hat{\phi}_i/\hat{\phi}_o)^{1/2}M^2 \tag{16}$$

and that Abbe's sine condition and Herschel's condition take their familiar forms.

We now show that image formation by electron lenses can be characterized with the aid of cardinal elements: foci, focal lengths, and principal planes. First, however, we must explain the novel notions of real and asymptotic imaging. So far, we have simply spoken of rotationally symmetric fields without specifying their distribution in space. Electron lenses are localized regions in which the magnetic or electrostatic field is strong and outside of which the field is weak but, in theory at least, does not vanish. Some typical lens geometries are shown in Fig. 2.

If the object and image are far from the lens, in effectively field-free space, or if the object is not a physical specimen but an intermediate image of the latter, the image formation can be analyzed in terms of the asymptotes to rays entering or emerging from the lens region. If, however, the true object or image is immersed within the lens field, as frequently occurs in the case of magnetic lenses, a different method of characterizing the lens properties must be adopted, and we shall speak of real cardinal elements. We consider the asymptotic case first.

It is convenient to introduce the solutions of Eq. (10) that satisfy the boundary conditions

$$\lim_{z\to-\infty} G(z) = 1; \qquad \lim_{z\to\infty} \bar{G}(z) = 1 \tag{17}$$

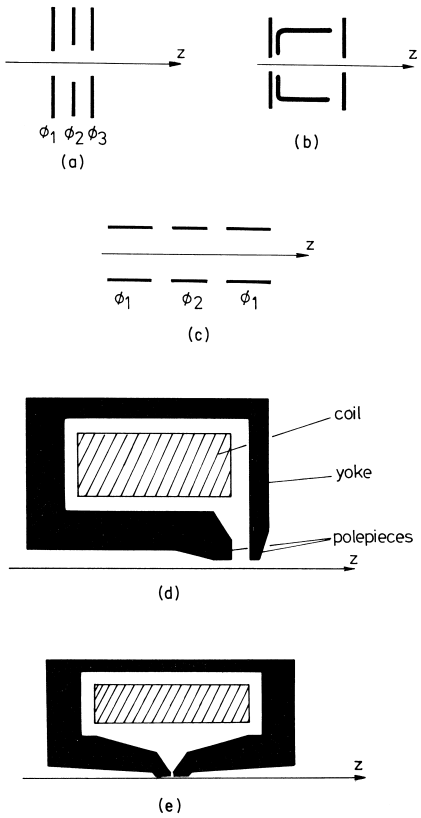


FIGURE 2 Typical electron lenses: (a–c) electrostatic lenses, of which (c) is an einzel lens; (d–e) magnetic lenses of traditional design.

These are rays that arrive at or leave the lens parallel to the axis (Fig. 3). As usual, the general solution is $x(z) = \alpha G(z) + \beta\bar{G}(z)$, where α and β are constants. We denote the emergent asymptote to $G(z)$ thus:

$$\lim_{z\to\infty} G(z) = G_i(z - z_{Fi}) \tag{18}$$

We denote the incident asymptote to $\bar{G}(z)$ thus:

$$\lim_{z\to-\infty} \bar{G}(z) = \bar{G}_o'(z - z_{Fo}) \tag{19}$$

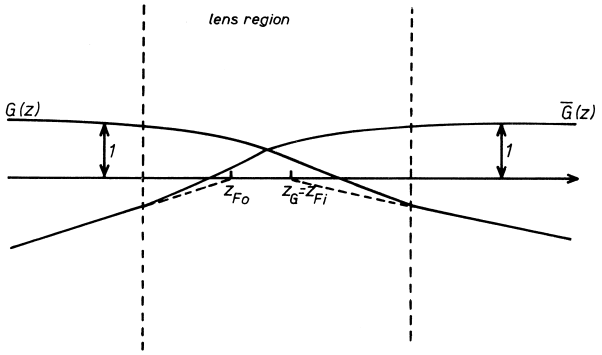


FIGURE 3 Rays $G(z)$ and $\bar{G}(z)$.

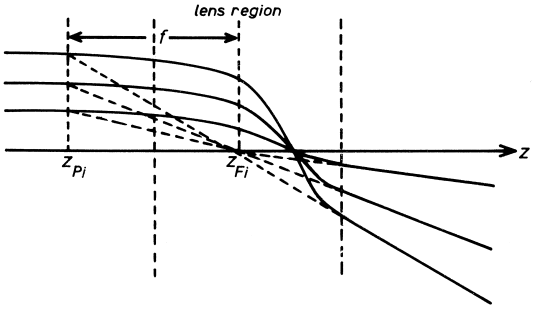


FIGURE 4 Focal and principal planes.

Clearly, all rays incident parallel to the axis have emergent asymptotes that intersect at $z = z_{\mathrm{Fi}}$; this point is known as the asymptotic image focus. It is not difficult to show that the emergent asymptotes to any family of rays that are parallel to one another but not to the axis intersect at a point in the plane $z = z_{\mathrm{Fi}}$. By applying a similar reasoning to $\bar{G}(z)$, we recognize that z_{Fo} is the asymptotic object focus. The incident and emergent asymptotes to $G(z)$ intersect in a plane z_{Pi}, which is known as the image principal plane (Fig. 4). The distance between z_{Fi} and z_{Pi} is the asymptotic image focal length:

$$z_{\mathrm{Fi}} - z_{\mathrm{Pi}} = -1/G'_{\mathrm{i}} = f_{\mathrm{i}} \tag{20}$$

We can likewise define z_{Po} and f_{o}:

$$z_{\mathrm{Po}} - z_{\mathrm{Fo}} = 1/\bar{G}'_{\mathrm{o}} = f_{\mathrm{o}} \tag{21}$$

The Wronskian tells us that $\hat{\phi}^{1/2}(G\bar{G}' - G'\bar{G})$ is constant and so

$$\hat{\phi}_{\mathrm{o}}^{1/2}\bar{G}'_{\mathrm{o}} = -\hat{\phi}_{\mathrm{i}}^{1/2}G'_{\mathrm{i}}$$

or

$$f_{\mathrm{o}}/\hat{\phi}_{\mathrm{o}}^{1/2} = f_{\mathrm{i}}/\hat{\phi}_{\mathrm{i}}^{1/2} \tag{22}$$

In magnetic lenses and electrostatic lenses, that provide no overall acceleration, $\hat{\phi}_{\mathrm{o}} = \hat{\phi}_{\mathrm{i}}$ and so $f_{\mathrm{o}} = f_{\mathrm{i}}$; we drop the subscript when no confusion can arise.

The coupling between an object space and an image space is conveniently expressed in terms of z_{Fo}, z_{Fi}, f_{o}, and f_{i}. From the general solution $x = \alpha G + \beta\bar{G}$, we see that

$$\begin{aligned} \lim_{z\to-\infty} x(z) &= \alpha + \beta(z - z_{\mathrm{Fo}})/f_{\mathrm{o}} \\ \lim_{z\to\infty} x(z) &= -\alpha(z - z_{\mathrm{Fi}})/f_{\mathrm{i}} + \beta \end{aligned} \tag{23}$$

and likewise for $y(z)$. Eliminating α and β, we find

$$\begin{bmatrix} x_2 \\ x'_2 \end{bmatrix} = \begin{bmatrix} -\dfrac{z_2 - z_{\mathrm{Fi}}}{f_{\mathrm{i}}} & f_{\mathrm{o}} + \dfrac{(z_1 - z_{\mathrm{Fo}})(z_2 - z_{\mathrm{Fi}})}{f_{\mathrm{i}}} \\ -\dfrac{1}{f_{\mathrm{i}}} & \dfrac{z_{\mathrm{o}} - z_{\mathrm{Fo}}}{f_{\mathrm{o}}} \end{bmatrix} \begin{bmatrix} x_1 \\ x'_1 \end{bmatrix} \tag{24}$$

where x_1 denotes $x(z)$ in some plane $z = z_1$ on the incident asymptote and x_2 denotes $x(z)$ in some plane $z = z_2$ on the emergent asymptote; $x' = dx/dz$. The matrix that appears in this equation is widely used to study systems with many focusing elements; it is known as the (paraxial) transfer matrix and takes slightly different forms for the various elements in use, quadrupoles in particular. We denote the transfer matrix by T.

If the planes z_1 and z_2 are conjugate, the point of arrival of a ray in z_2 will vary with the position coordinates of its point of departure in z_1 but will be independent of the gradient at that point. The transfer matrix element T_{12} must therefore vanish,

$$(z_{\mathrm{o}} - z_{\mathrm{Fo}})(z_{\mathrm{i}} - z_{\mathrm{Fi}}) = -f_{\mathrm{o}}f_{\mathrm{i}} \tag{25}$$

in which we have replaced z_1 and z_2 by z_{o} and z_{i} to indicate that these are now conjugates (object and image). This is the familiar lens equation in Newtonian form. Writing $z_{\mathrm{Fi}} = z_{\mathrm{Pi}} + f_{\mathrm{i}}$ and $z_{\mathrm{Fo}} = z_{\mathrm{Po}} - f_{\mathrm{o}}$, we obtain

$$\frac{f_{\mathrm{o}}}{z_{\mathrm{Po}} - z_{\mathrm{o}}} + \frac{f_{\mathrm{i}}}{z_{\mathrm{i}} - z_{\mathrm{Pi}}} = 1 \tag{26}$$

the thick-lens form of the regular lens equation.

Between conjugates, the matrix T takes the form

$$T = \begin{bmatrix} M & 0 \\ -\dfrac{1}{f_{\mathrm{i}}} & \dfrac{f_{\mathrm{o}}}{f_{\mathrm{i}}}\dfrac{1}{M} \end{bmatrix} \tag{27}$$

in which M denotes the asymptotic magnification, the height of the image asymptote to $G(z)$ in the image plane.

If, however, the object is a real physical specimen and not a mere intermediate image, the asymptotic cardinal elements cannot in general be used, because the object may well be situated inside the field region and only a part of the field will then contribute to the image formation. Fortunately, objective lenses, in which this situation arises, are normally operated at high magnification with the specimen close to the real object focus, the point at which the ray $\bar{G}(z)$ itself intersects the axis [whereas the asymptotic object focus is the point at which the asymptote to $\bar{G}(z)$ in object space intersects the optic axis]. The corresponding real focal length is then defined by the slope of $\bar{G}(z)$ at the object focus F_{o}: $f = 1/G'(F_{\mathrm{o}})$; see Fig. 5.

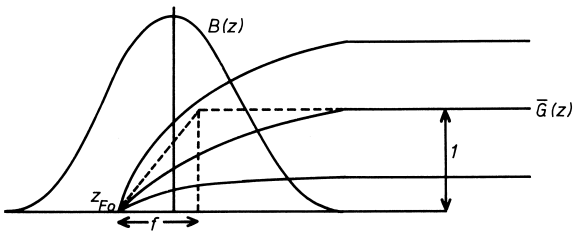


FIGURE 5 Real focus and focal length.

2. Quadrupoles

In the foregoing discussion, we have considered only rotationally symmetric fields and have needed only the axial distributions $B(z)$ and $\phi(z)$. The other symmetry of greatest practical interest is that associated with electrostatic and magnetic quadrupoles, widely used in particle accelerators. Here, the symmetry is lower, the fields possessing planes of symmetry and antisymmetry only; these planes intersect in the optic axis, and we shall assume forthwith that electrostatic and magnetic quadrupoles are disposed as shown in Fig. 6. The reason for this is simple: The paraxial equations of motion for charged particles traveling through quadrupoles separate into two uncoupled equations only if this choice is adopted. This is not merely a question of mathematical convenience; if quadrupole fields overlap and the total system does not have the symmetry indicated, the desired imaging will not be achieved.

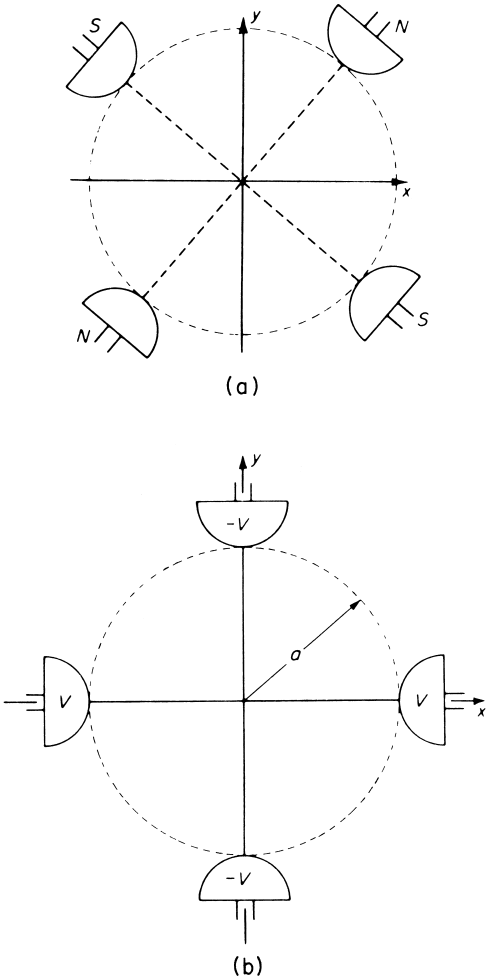


FIGURE 6 (a) Magnetic and (b) electrostatic quadrupoles.

The paraxial equations are now different in the x–z and y–z planes:

$$\begin{aligned}\frac{d}{dz}(\hat{\phi}^{1/2}x') + \frac{\gamma\phi'' - 2\gamma p_2 + 4\eta Q_2\hat{\phi}^{1/2}}{4\hat{\phi}^{1/2}}x = 0\\ \frac{d}{dz}(\hat{\phi}^{1/2}y') + \frac{\gamma\phi'' + 2\gamma p_2 - 4\eta Q_2\hat{\phi}^{1/2}}{4\hat{\phi}^{1/2}}y = 0\end{aligned} \tag{28}$$

in which we have retained the possible presence of a round electrostatic lens field $\phi(z)$. The functions $p_2(z)$ and $Q_2(z)$ that also appear characterize the quadrupole fields; their meaning is easily seen from the field expansions [for $B(z) = 0$]:

$$\begin{aligned}\Phi(x, y, z) = \phi(z) &- \frac{1}{4}(x^2 + y^2)\phi''(z)\\ &+ \frac{1}{64}(x^2 + y^2)^2\phi^{(4)}(z)\\ &+ \frac{1}{2}(x^2 - y^2)p_2(z) - \frac{1}{24}(x^4 + y^4)p_2''(z)\\ &+ \frac{1}{24}p_4(z)(x^4 - 6x^2y^2 + y^4) + \cdots\\ \Phi(r, \psi, z) = \phi(z) &- \frac{1}{4}r^2\phi'' + \frac{1}{64}r^4\phi^{(4)}\\ &+ \frac{1}{2}p_2r^2\cos 2\psi - \frac{1}{24}p_2''r^4\cos 2\psi\\ &+ \frac{1}{24}p_4r^4\cos 4\psi + \cdots\end{aligned} \tag{29}$$

$$\begin{aligned}A_x &= -\frac{x}{12}(x^2 - 3y^2)Q_2'(z)\\ A_y &= \frac{y}{12}(y^2 - 3x^2)Q_2'(z)\\ A_z &= \frac{1}{2}(x^2 - y^2)Q_2(z) - \frac{1}{24}(x^4 - y^4)Q_2''(z)\\ &\quad + \frac{1}{24}(x^4 - 6x^2y^2 + y^4)Q_4(z)\end{aligned} \tag{30}$$

The terms $p_4(z)$ and $Q_4(z)$ characterize octopole fields, and we shall refer to them briefly in connection with the aberration correction below.

It is now necessary to define separate transfer matrices for the x–z plane and for the y–z plane. These have exactly the same form as Eqs. (24) and (27), but we have to distinguish between two sets of cardinal elements. For arbitrary planes z_1 and z_2, we have

$$T^{(x)} = \begin{bmatrix} -\dfrac{z_2 - z_{\mathrm{Fi}}^{(x)}}{f_{x\mathrm{i}}} & \dfrac{\left(z_2 - z_{\mathrm{Fi}}^{(x)}\right)\left(z_2 - z_{\mathrm{Fo}}^{(x)}\right)}{f_{x\mathrm{o}}} + f_{x\mathrm{i}} \\ -\dfrac{1}{f_{x\mathrm{i}}} & \dfrac{z_1 - z_{\mathrm{Fo}}^{(x)}}{f_{x\mathrm{i}}} \end{bmatrix}$$

$$T^{(y)} = \begin{bmatrix} -\dfrac{z_2 - z_{\mathrm{Fi}}^{(y)}}{f_{y\mathrm{i}}} & \dfrac{\left(z_2 - z_{\mathrm{Fi}}^{(y)}\right)\left(z_1 - z_{\mathrm{Fo}}^{(y)}\right)}{f_{y\mathrm{o}}} + f_{y\mathrm{i}} \\ -\dfrac{1}{f_{y\mathrm{i}}} & \dfrac{z_1 - z_{\mathrm{Fo}}^{(y)}}{f_{y\mathrm{i}}} \end{bmatrix}. \tag{31}$$

Suppose now that $z = z_{x\mathrm{o}}$ and $z = z_{x\mathrm{i}}$ and conjugate so that $T_{12}^{(x)} = 0$; in general, $T_{12}^{(y)} \neq 0$ and so a point in the object plane $z = z_{x\mathrm{o}}$ will be imaged as a line parallel to the y axis. Similarly, if we consider a pair of conjugates $z = z_{y\mathrm{o}}$ and $z = z_{y\mathrm{i}}$, we obtain a line parallel to the x axis. The imaging is hence astigmatic, and the astigmatic differences in object and image space can be related to the magnification

$$\begin{aligned} \wedge_{\mathrm{i}} &:= z_{x\mathrm{i}} - z_{y\mathrm{i}} = \wedge_{\mathrm{Fi}} - f_{x\mathrm{i}} M_x + f_{y\mathrm{i}} M_y \\ \wedge_{\mathrm{i}} &:= z_{x\mathrm{o}} - z_{y\mathrm{o}} = \wedge_{\mathrm{Fo}} + f_{x\mathrm{o}}/M_x - f_{y\mathrm{o}}/M_y, \end{aligned} \tag{32}$$

where

$$\begin{aligned} \wedge_{\mathrm{Fi}} &:= z_{\mathrm{Fi}}^{(x)} - z_{\mathrm{Fi}}^{(y)} = \wedge_{\mathrm{i}}(M_x = M_y = 0) \\ \wedge_{\mathrm{Fo}} &:= z_{\mathrm{Fo}}^{(x)} - z_{\mathrm{Fo}}^{(y)} = \wedge_{\mathrm{o}}(M_x = M_y \to \infty). \end{aligned} \tag{33}$$

Solving the equations $\wedge_{\mathrm{i}} = \wedge_{\mathrm{o}} = 0$ for M_x and M_y, we find that there is a pair of object planes for which the image is stigmatic though not free of distortion.

3. Prisms

There is an important class of devices in which the optic axis is not straight but a simple curve, almost invariably lying in a plane. The particles remain in the vicinity of this curve, but they experience different focusing forces in the plane and perpendicular to it. In many cases, the axis is a circular arc terminated by straight lines. We consider the situation in which charged particles travel through a magnetic sector field (Fig. 7); for simplicity, we assume that the field falls abruptly to zero at entrance and exit planes (rather than curved surfaces) and that the latter are normal to the optic axis, which is circular. We regard the plane containing the axis as horizontal. The vertical field at the axis is denoted by B_{o}, and off the axis, $B = B_{\mathrm{o}}(r/R)^{-n}$ in the horizontal plane. It can then be shown, with the notation of Fig. 7, that paraxial trajectory equations of the form

$$x'' + k_{\mathrm{v}}^2 x = 0; \qquad y'' + k_{\mathrm{H}}^2 y = 0 \tag{34}$$

describe the particle motion, with $k_{\mathrm{H}}^2 = (1-n)/R^2$ and $k_{\mathrm{v}}^2 = n/R^2$. Since these are identical in appearance with the quadrupole equations but do not have different signs, the particles will be focused in both directions but not in the same "image" plane unless $k_{\mathrm{H}} = k_{\mathrm{v}}$ and hence $n = \frac{1}{2}$. The cases $n = 0$, for which the magnetic field is homogeneous, and $n = \frac{1}{2}$ have been extensively studied. Since prisms are widely used to separate particles of different energy or momentum, the dispersion is an important quantity, and the transfer matrices are usually extended to include this information. In practice, more complex end faces are employed than the simple planes normal to the axis considered here, and the fringing fields cannot be completely neglected, as they are in the sharp cutoff approximation.

Electrostatic prisms can be analyzed in a similar way and will not be discussed separately.

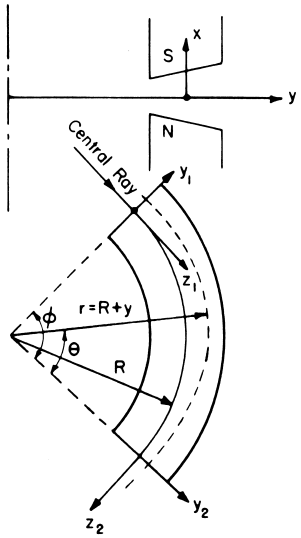


FIGURE 7 Passage through a sector magnet.

B. Aberrations

1. Traditional Method

The paraxial approximation describes the dominant focusing in the principal electron-optical devices, but this is inevitably perturbed by higher order effects, or aberrations. There are several kinds of aberrations. By retaining higher order terms in the field or potential expansions, we obtain the family of geometric aberrations. By considering small changes in particle energy and lens strength, we obtain the chromatic aberrations. Finally, by examining the effect of small departures from the assumed symmetry of the field, we obtain the parasitic aberrations.

All these types of aberrations are conveniently studied by means of perturbation theory. Suppose that we have obtained the paraxial equations as the Euler–Lagrange equations of the paraxial form of M [Eq. (6)], which we denote

$M^{(\mathrm{P})}$. Apart from a trivial change of scale, we have

$$M^{(\mathrm{P})} = -(1/8\hat{\phi}^{1/2})(\gamma\phi'' + \eta^2 B^2)(x^2 + y^2) + \frac{1}{2}\hat{\phi}^{1/2}(x'^2 + y'^2) \tag{35}$$

Suppose now that $M^{(\mathrm{P})}$ is perturbed to $M^{(\mathrm{P})} + M^{(\mathrm{A})}$. The second term $M^{(\mathrm{A})}$ may represent additional terms, neglected in the paraxial approximation, and will then enable us to calculate the geometric aberrations; alternatively, $M^{(\mathrm{A})}$ may measure the change in $M^{(\mathrm{P})}$ when particle energy and lens strength fluctuate, in which case it tells us the chromatic aberration. Other field terms yield the parasitic aberration. We illustrate the use of perturbation theory by considering the geometric aberrations of round lenses. Here, we have

$$\begin{aligned} M^{(\mathrm{A})} = M^{(4)} = &-\frac{1}{4}L_1(x^2 + y^2)^2 \\ &- \frac{1}{2}L_2(x^2 + y^2)(x'^2 + y'^2) \\ &- \frac{1}{4}L_3(x'^2 + y'^2)^2 \\ &- R(xy' - x'y)^2 \\ &- P\hat{\phi}^{1/2}(x^2 + y^2)(xy' - x'y) \\ &- Q\hat{\phi}^{1/2}(x'^2 + y'^2)(xy' - x'y) \end{aligned} \tag{36}$$

with

$$\begin{aligned} L_1 &= \frac{1}{32\hat{\phi}^{1/2}}\left(\frac{\phi''^2}{\hat{\phi}} - \gamma\phi^{(4)} + \frac{2\gamma\phi''\eta^2 B^2}{\hat{\phi}} + \frac{\eta^4 B^4}{\hat{\phi}} - 4\eta^2 B B''\right) \\ L_2 &= \frac{1}{8\hat{\phi}^{1/2}}(\gamma\phi'' + \eta^2 B^2) \\ L_3 &= \frac{1}{2}\hat{\phi}^{1/2}; \quad P = \frac{\eta}{16\hat{\phi}^{1/2}}\left(\frac{\gamma\phi'' B}{\hat{\phi}} + \frac{\eta^2 B^2}{\hat{\phi}} - B''\right) \\ Q &= \frac{\eta B}{4\hat{\phi}^{1/2}}; \quad R = \frac{\eta^2 B^2}{8\hat{\phi}^{1/2}} \end{aligned} \tag{37}$$

and with $S^{(\mathrm{A})} = \int_{z_0}^{z} M^{(4)}\,dz$, we can show that

$$\begin{aligned} \frac{\partial S^{(\mathrm{A})}}{\partial x_\mathrm{a}} &= p_x^{(\mathrm{A})} t(z) - x^{(\mathrm{A})}\hat{\phi}^{1/2} t'(z) \\ \frac{\partial S^{(\mathrm{A})}}{\partial y_\mathrm{a}} &= p_x^{(\mathrm{A})} s(z) - x^{(\mathrm{A})}\hat{\phi}^{1/2} s'(z) \end{aligned} \tag{38}$$

where $s(z)$ and $t(z)$ are the solutions of Eq. (10) for which $s(z_0) = t(z_\mathrm{a}) = 1$, $s(z_\mathrm{a}) = t(z_0) = 0$, and $z = z_\mathrm{a}$ denotes some aperture plane. Thus, in the image plane,

$$x^{(\mathrm{A})} = -(M/W)\partial S_{\mathrm{oi}}^{(\mathrm{A})}/\partial x_\mathrm{a} \tag{39}$$

where $S_{\mathrm{oi}}^{(\mathrm{A})}$ denotes $\int_{z_\mathrm{o}}^{z_\mathrm{i}} M^{(4)}\,dz$, with a similar expression for $y^{(\mathrm{A})}$. The quantities with superscript (A) indicate the departure from the paraxial approximation, and we write

$$\begin{aligned} \Delta x_\mathrm{i} &= x^{(\mathrm{A})}/M = -(1/W)\,\partial S_{\mathrm{oi}}^{(\mathrm{A})}/\partial x_\mathrm{a} \\ \Delta y_\mathrm{i} &= y^{(A)}/M = -(1/W)\,\partial S_{\mathrm{oi}}^{(\mathrm{A})}/\partial y_{(\mathrm{a})} \end{aligned} \tag{40}$$

The remainder of the calculation is lengthy but straightforward. Into $M^{(4)}$, the paraxial solutions are substituted and the resulting terms are grouped according to their dependence on x_o, y_o, x_a, and y_a. We find that $S^{(\mathrm{A})}$ can be written

$$\begin{aligned} -S^{(\mathrm{A})}/W = &\frac{1}{4}Er_\mathrm{o}^4 + \frac{1}{4}Cr_\mathrm{a}^4 + \frac{1}{2}A(V^2 - v^2) \\ &+ \frac{1}{2}Fr_\mathrm{o}^2 r_\mathrm{a}^2 + Dr_\mathrm{o}^2 V + Kr_\mathrm{a}^2 V \\ &+ v\left(dr_\mathrm{o}^2 + kr_\mathrm{a}^2 + aV\right) \end{aligned} \tag{41}$$

with

$$\begin{aligned} r_\mathrm{o}^2 &= x_\mathrm{o}^2 + y_\mathrm{o}^2; & r_\mathrm{a}^2 &= x_a^2 + y_a^2 \\ V &= x_\mathrm{o}x_\mathrm{a} + y_\mathrm{o}y_\mathrm{a}; & v &= x_\mathrm{o}y_\mathrm{a} - x_\mathrm{a}y_\mathrm{o} \end{aligned} \tag{42}$$

and

$$\begin{aligned} \Delta x_\mathrm{i} = &x_\mathrm{a}\left[Cr_\mathrm{a}^2 + 2KV + 2kv + (F - A)r_\mathrm{o}^2\right] \\ &+ x_\mathrm{o}\left(Kr_\mathrm{a}^2 + 2AV + av + Dr_\mathrm{o}^2\right) \\ &- y_\mathrm{o}\left(kr_\mathrm{a}^2 + aV + dr_\mathrm{o}^2\right) \\ \Delta y_\mathrm{i} = &y_\mathrm{a}\left[Cr_\mathrm{a}^2 + 2KV + 2kv + (F - A)r_\mathrm{o}^2\right] \\ &+ x_\mathrm{o}\left(kr_\mathrm{a}^2 + aV + dr_\mathrm{o}^2\right) \end{aligned} \tag{43}$$

Each coefficient A, $C, \ldots, d$, k represents a different type of geometric aberration. Although all lenses suffer from every aberration, with the exception of the anisotropic aberrations described by k, a, and d, which are peculiar to magnetic lenses, the various aberrations are of very unequal importance when lenses are used for different purposes. In microscope objectives, for example, the incident electrons are scattered within the specimen and emerge at relatively steep angles to the optic axis (several milliradians or tens of milliradians). Here, it is the spherical (or aperture) aberration C that dominates, and since this aberration does not vanish on the optic axis, being independent of r_o, it has an extremely important effect on image quality. Of the geometric aberrations, it is this spherical aberration that determines the resolving power of the electron microscope. In the subsequent lenses of such instruments, the image is progressively enlarged until the final magnification, which may reach 100,000× or 1,000,000×, is attained. Since angular magnification is inversely proportional to transverse magnification, the

angular spread of the beam in these projector lenses will be tiny, whereas the off-axis distance becomes large. Here, therefore, the distortions D and d are dominant.

A characteristic aberration figure is associated with each aberration. This figure is the pattern in the image plane formed by rays from some object point that cross the aperture plane around a circle. For the spherical aberration, this figure is itself a circle, irrespective of the object position, and the effect of this aberration is therefore to blur the image uniformly, each Gaussian image point being replaced by a disk of radius MCr_a^3. The next most important aberration for objective lenses is the coma, characterized by K and k, which generates the comet-shaped streak from which it takes its name. The coefficients A and F describe Seidel astigmatism and field curvature, respectively; the astigmatism replaces stigmatic imagery by line imagery, two line foci being formed on either side of the Gaussian image plane, while the field curvature causes the image to be formed not on a plane but on a curved image surface. The distortions are more graphically understood by considering their effect on a square grid in the object plane. Such a grid is swollen or shrunk by the isotropic distortion D and warped by the anisotropic distortion d; the latter has been evocatively styled as a pocket handkerchief distortion. Figure 8 illustrates these various aberrations.

Each aberration has a large literature, and we confine this account to the spherical aberration, an eternal preoccupation of microscope lens designers. In practice, it is more convenient to define this in terms of angle at the specimen, and recalling that $x(z) = x_o s(z) + x_a t(z)$, we see that $x'_o = x_o s'(z_o) + x_a t'(z_o)$ Hence,

$$\Delta x_i = C x_a \left(x_a^2 + y_a^2\right) = \frac{C}{t_o'^3} x'_o \left(x_o'^2 + y_o'^2\right) + \cdots \quad (44)$$

and we therefore write $C_s = c/t_o'^3$ so that

$$\Delta x_i = C_s x'_o \left(x_o'^2 + y_o'^2\right); \qquad y_i = C_s y'_o \left(x_o'^2 + y_o'^2\right) \quad (45)$$

It is this coefficient C_s that is conventionally quoted and tabulated. A very important and disappointing property of C_s is that it is intrinsically positive: The formula for it can be cast into positive-definite form, which means that we cannot hope to design a round lens free of this aberration by skillful choice of geometry and excitation. This result is known as Scherzer's theorem. An interesting attempt to upset the theorem was made by Glaser, who tried setting the integrand that occurs in the formula for C_s, and that can be written as the sum of several squared terms, equal to zero and solving the resulting differential equation for the field (in the magnetic case). Alas, the field distribution that emerged was not suitable for image formation, thus confirming the truth of the theorem, but it has been found useful in β-ray spectroscopy. The full implications of the theorem were established by Werner Tretner, who established the lower limit for C_s as a function of the practical constraints imposed by electrical breakdown, magnetic saturation, and geometry.

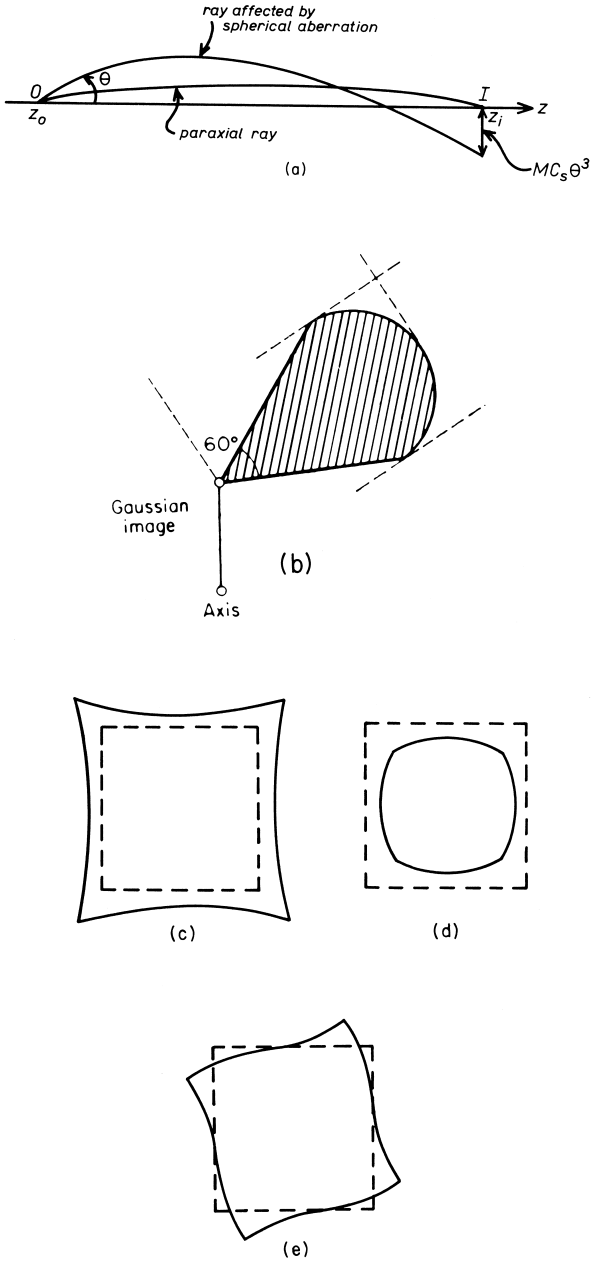


FIGURE 8 Aberration patterns: (a) spherical aberration; (b) coma; (c–e) distortions.

Like the cardinal elements, the aberrations of objective lenses require a slightly different treatment from those of condenser lenses and projector lenses. The reason is easily understood: In magnetic objective lenses (and probe-forming lenses), the specimen (or target) is commonly immersed deep inside the field and only the field region downstream contributes to the image formation. The

spherical aberration is likewise generated only by this part of the field, and the expression for C_s as an integral from object plane to image plane reflects this. In other lenses, however, the object is in fact an intermediate image, formed by the next lens upstream, and the whole lens field contributes to the image formation and hence to the aberrations. It is then the coupling between incident and emergent asymptotes that is of interest, and the aberrations are characterized by asymptotic aberration coefficients. These exhibit an interesting property: They can be expressed as polynomials in reciprocal magnification m ($m = 1/M$), with the coefficients in these polynomials being determined by the lens geometry and excitation and independent of magnification (and hence of object position). This dependence can be written

$$\begin{bmatrix} C \\ K \\ A \\ F \\ D \end{bmatrix} = Q \begin{bmatrix} m^4 \\ m^3 \\ m^2 \\ m \\ 1 \end{bmatrix} \begin{bmatrix} k \\ a \\ d \end{bmatrix} = q \begin{bmatrix} m^2 \\ m \\ 1 \end{bmatrix} \tag{46}$$

in which Q and q have the patterns

$$Q = \begin{bmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \end{bmatrix}; \qquad q = \begin{bmatrix} x & x & x \\ 0 & x & x \\ 0 & 0 & x \end{bmatrix}, \tag{47}$$

where an x indicates that the matrix element is a nonzero quantity determined by the lens geometry and excitation.

Turning now to chromatic aberrations, we have

$$m^{(P)} = \frac{\partial m^{(2)}}{\partial \phi}\Delta\phi + \frac{\partial m^{(2)}}{\partial B}\Delta B \tag{48}$$

and a straightforward calculation yields

$$\begin{aligned} \Delta x^{(c)} &= -(C_c x_o' + C_D x_o - C_\theta y_o)\left(\frac{\gamma\Delta\phi_o}{\hat{\phi}_o} - 2\frac{\Delta B_0}{B_0}\right) \\ \Delta y^{(c)} &= -(C_c y_o + C_D y_o + C_\theta x_o)\left(\frac{\gamma\Delta\phi_o}{\hat{\phi}_o} - 2\frac{\Delta B_0}{B_0}\right) \end{aligned} \tag{49}$$

for magnetic lenses or

$$x^{(c)} = -(C_c x_o' + C_D x_o)\frac{\Delta\phi_o}{\hat{\phi}_o} \tag{50}$$

with a similar expression for $y^{(c)}$ for electrostatic lenses. In objective lenses, the dominant aberration is the (axial) chromatic aberration C_c, which causes a blur in the image that is independent of the position of the object point, like that due to C_s. The coefficient C_c also shares with C_s the property of being intrinsically positive. The coefficients C_D and C_θ affect projector lenses, but although they are pure distortions, they may well cause blurring since the term in $\Delta\phi_o$ and ΔB_o represents a spread, as in the case of the initial electron energy, or an oscillation typically at main frequency, coming from the power supplies.

Although a general theory can be established for the parasitic aberrations, this is much less useful than the theory of the geometric and chromatic aberrations; because the parasitic aberrations are those caused by accidental, unsystematic errors—imperfect roundness of the openings in a round lens, for example, or inhomogeneity of the magnetic material of the yoke of a magnetic lens, or imperfect alignment of the polepieces or electrodes. We therefore merely point out that one of the most important parasitic aberrations is an axial astigmatism due to the weak quadrupole field component associated with ellipticity of the openings. So large is this aberration, even in carefully machined lenses, that microscopes are equipped with a variable weak quadrupole, known as a stigmator, to cancel this unwelcome effect.

We will not give details of the aberrations of quadrupoles and prisms here. Quadrupoles have more independent aberrations than round lenses, as their lower symmetry leads us to expect, but these aberrations can be grouped into the same families: aperture aberrations, comas, field curvatures, astigmatisms, and distortions. Since the optic axis is straight, they are third-order aberrations, like those of round lenses, in the sense that the degree of the dependence on x_o, x_o', y_o, and y_o' is three. The primary aberrations of prisms, on the other hand, are of second order, with the axis now being curved.

2. Lie Methods

An alternative way of using Hamiltonian mechanics to study the motion of charged particles has been developed, by Alex Dragt and colleagues especially, in which the properties of Lie algebra are exploited. This has come to be known as Lie optics. It has two attractions, one very important for particle optics at high energies (accelerator optics): first, interrelations between aberration coefficients are easy to establish, and second, high-order perturbations can be studied systematically with the aid of computer algebra and, in particular, of the differential algebra developed for the purpose by Martin Berz. At lower energies, the Lie methods provide a useful check of results obtained by the traditional procedures, but at higher energies they give valuable information that would be difficult to obtain in any other way.

C. Instrumental Optics: Components

1. Guns

The range of types of particle sources is very wide, from the simple triode gun with a hairpin-shaped filament relying on thermionic emission to the plasma sources furnishing high-current ion beams. We confine this account to the thermionic and field-emission guns that are used in electron-optical instruments to furnish modest electron currents: thermionic guns with tungsten or lanthanum hexaboride emitters, in which the electron emission is caused by heating the filament, and field-emission guns, in which a very high electric field is applied to a sharply pointed tip (which may also be heated). The current provided by the gun is not the only parameter of interest and is indeed often not the most crucial. For microscope applications, a knowledge of brightness B is much more important; this quantity is a measure of the quality of the beam. Its exact definition requires considerable care, but for our present purposes it is sufficient to say that it is a measure of the current density per unit solid angle in the beam. For a given current, the brightness will be high for a small area of emission and if the emission is confined to a narrow solid angle. In scanning devices, the writing speed and the brightness are interrelated, and the resulting limitation is so severe that the scanning transmission electron microscope (STEM) came into being only with the development of high-brightness field-emission guns. Apart from a change of scale with $\hat{\phi}_2/\hat{\phi}_1$ in accelerating structures, the brightness is a conserved quantity in electron-optical systems (provided that the appropriate definition of brightness is employed).

The simplest and still the most widely used electron gun is the triode gun, consisting of a heated filament or cathode, an anode held at a high positive potential relative to the cathode, and, between the two, a control electrode known as the *wehnelt*. The latter is held at a small negative potential relative to the cathode and serves to define the area of the cathode from which electrons are emitted. The electrons converge to a waist, known as the crossover, which is frequently within the gun itself (Fig. 9). If j_c is the current density at the center of this crossover and α_s is the angular spread (defined in Fig. 9), then

$$B = j_c/\pi\alpha_s^2 \tag{51}$$

It can be shown that B cannot exceed the Langmuir limit $B_{max} = je\phi/\pi kT$, in which j is the current density at the filament, ϕ is the accelerating voltage, k is Boltzmann's constant (1.4×10^{-23} J/K), and T is the filament temperature. The various properties of the gun vary considerably with the size and position of the wehnelt and anode and the potentials applied to them; the general behavior has been satisfactorily explained in terms of a rather simple model by Rolf Lauer.

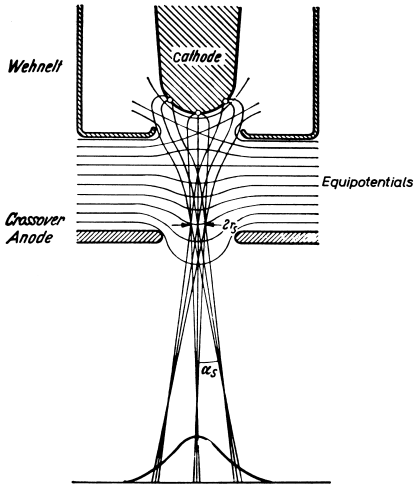


FIGURE 9 Electron gun and formation of the crossover.

The crossover is a region in which the current density is high, and frequently high enough for interactions between the beam electrons to be appreciable. A consequence of this is a redistribution of the energy of the particles and, in particular, an increase in the energy spread by a few electron volts. This effect, detected by Hans Boersch in 1954 and named after him, can be understood by estimating the mean interaction using statistical techniques.

Another family of thermionic guns has rare-earth boride cathodes, LaB_6 in particular. These guns were introduced in an attempt to obtain higher brightness than a traditional thermionic gun could provide, and they are indeed brighter sources; they are technologically somewhat more complex, however. They require a slightly better vacuum than tungsten triode guns, and in the first designs the LaB_6 rod was heated indirectly by placing a small heating coil around it; subsequently, however, directly heated designs were developed, which made these guns more attractive for commercial purposes.

Even LaB_6 guns are not bright enough for the needs of the high-resolution STEM, in which a probe only a few tenths of a nanometer in diameter is raster-scanned over a thin specimen and the transmitted beam is used to form an image (or images). Here, a field-emission gun is indispensable. Such guns consist of a fine tip and two (or more) electrodes, the first of which creates a very high electric field at the tip, while the second accelerates the electrons thus extracted to the desired accelerating voltage. Such guns operate satisfactorily only if the vacuum is very good indeed; the pressure in a field-emission gun must be some five or six orders of magnitude higher than that in a thermionic triode gun. The resulting brightness is

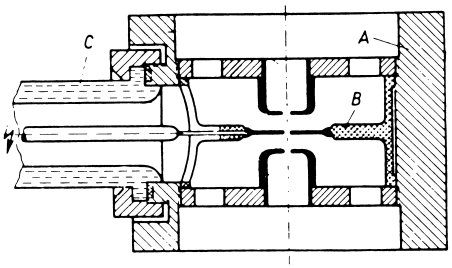


FIGURE 10 Electrostatic einzel lens design: (A) lens casing; (B and C) insulators.

appreciably higher, but the current is not always sufficient when only a modest magnification is required.

We repeat that the guns described above form only one end of the spectrum of particle sources. Others have large flat cathodes. Many are required to produce high currents and current densities, in which case we speak of space-charge flow; these are the Pierce guns and PIGs (Pierce ion guns).

2. Electrostatic Lenses

Round electrostatic lenses take the form of a series of plates in which a round opening has been pierced or circular cylinders all centered on a common axis (Fig. 10). The potentials applied may be all different or, more often, form a simple pattern. The most useful distinction in practice separates lenses that create no overall acceleration of the beam (although, of course, the particles are accelerated and decelerated within the lens field) and those that do produce an overall acceleration or deceleration. In the first case, the usual configuration is the einzel lens, in which the outer two of the three electrodes are held at anode potential (or at the potential of the last electrode of any lens upstream if this is not at anode potential) and the central electrode is held at a different potential. Such lenses were once used in electrostatic microscopes and are still routinely employed when the insensitivity of electrostatic systems to voltage fluctuations that affect all the potentials equally is exploited. Extensive sets of curves and tables describing the properties of such lenses are available.

Accelerating lenses with only a few electrodes have also been thoroughly studied; a configuration that is of interest today is the multielectrode accelerator structure. These accelerators are not intended to furnish very high particle energies, for which very different types of accelerator are employed, but rather to accelerate electrons to energies beyond the limit of the simple triode structure, which cannot be operated above ~150 kV. For microscope and microprobe instruments with accelerating voltages in the range of a few hundred kilovolts up to a few megavolts, therefore, an accelerating structure must be inserted between the gun and the first condenser lens. This structure is essentially a multielectrode electrostatic lens with the desired accelerating voltage between its terminal electrodes. This point of view is particularly useful when a field-emission gun is employed because of an inconvenient aspect of the optics of such guns: The position and size of the crossover vary with the current emitted. In a thermionic gun, the current is governed essentially by the temperature of the filament and can hence be varied by changing the heating current. In field-emission guns, however, the current is determined by the field at the tip and is hence varied by changing the potential applied to the first electrode, which in turn affects the focusing field inside the gun. When such a gun is followed by an accelerator, it is not easy to achieve a satisfactory match for all emission currents and final accelerating voltages unless both gun and accelerator are treated as optical elements. Miniature lenses and guns and arrays of these are being fabricated, largely to satisfy the needs of nanolithography. A spectacular achievement is the construction of a scanning electron microscope that fits into the hand, no bigger than a pen. The optical principles are the same as for any other lens.

3. Magnetic Lenses

There are several kinds of magnetic lenses, but the vast majority have the form of an electromagnet pierced by a circular canal along which the electrons pass. Figure 11 shows such a lens schematically, and Fig. 12 illustrates a more realistic design in some detail. The magnetic flux

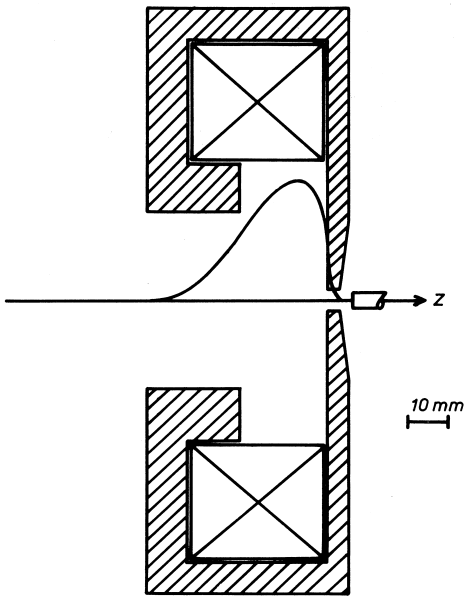


FIGURE 11 Typical field distribution in a magnetic lens.

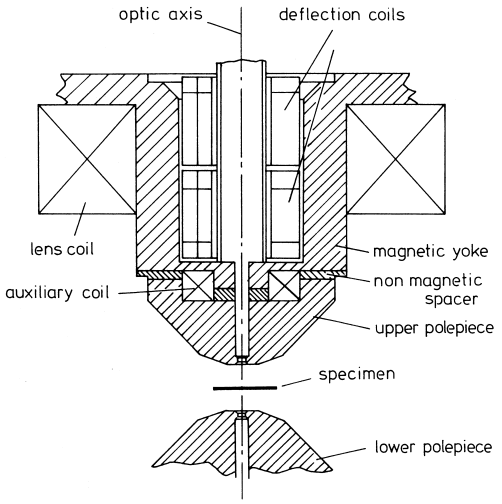


FIGURE 12 Modern magnetic objective lens design. (Courtesy of Philips, Eindhoven.)

is provided by a coil, which usually carries a low current through a large number of turns; water cooling prevents overheating. The magnetic flux is channeled through an iron yoke and escapes only at the gap, where the yoke is terminated with polepieces of high permeability. This arrangement is chosen because the lens properties will be most favorable if the axial magnetic field is in the form of a high, narrow bell shape (Fig. 11) and the use of a high-permeability alloy at the polepieces enables one to create a strong axial field without saturating the yoke. Considerable care is needed in designing the exact shape of these polepieces, but for a satisfactory choice, the properties of the lens are essentially determined by the gap S, the bore D (or the front and back bores if these are not the same), and the excitation parameter J; the latter is defined by $J = NI/\hat{\phi}_o^{1/2}$, where NI is the product of the number of turns of the coil and the current carried by it and $\hat{\phi}_o$ is the relativistic accelerating voltage; S and D are typically of the order of millimeters and J is a few amperes per (volts)$^{1/2}$. The quantity NI can be related to the axial field strength with the aid of Ampère's circuital theorem (Fig. 13); we see that

$$\int_{-\infty}^{\infty} B(z)\,dz = \mu_0 NI \qquad \text{so that} \quad NI \propto B_0$$

the maximum field in the gap, the constant of proportionality being determined by the area under the normalized flux distribution $B(z)/B_0$.

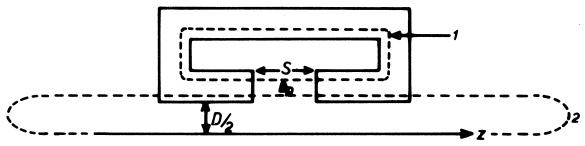


FIGURE 13 Use of Ampère's circuital theorem to relate lens excitation to axial field strength.

Although accurate values of the optical properties of magnetic lenses can be obtained only by numerical methods, in which the field distribution is first calculated by one of the various techniques available—finite differences, finite elements, and boundary elements in particular—their variation can be studied with the aid of field models. The most useful (though not the most accurate) of these is Glaser's bell-shaped model, which has the merits of simplicity, reasonable accuracy, and, above all, the possibility of expressing all the optical quantities such as focal length, focal distance, the spherical and chromatic aberration coefficients C_s and C_c, and indeed all the third-order aberration coefficients, in closed form, in terms of circular functions. In this model, $B_{(z)}$ is represented by

$$B_{(z)} = B_0/(1 + z^2/a^2) \tag{52}$$

and writing $w^2 = 1 + k^2$, $k^2 = \eta^2 B_0^2 a^2/4\hat{\phi}_0$, $z = a\cot\psi$ the paraxial equation has the general solution

$$x(\psi) = (A\cos\psi + B\sin\psi)/\sin\psi \tag{53}$$

The focal length and focal distance can be written down immediately, and the integrals that give C_s and C_c can be evaluated explicitly. This model explains very satisfactorily the way in which these quantities vary with the excitation and with the geometric parameter a.

The traditional design of Fig. 12 has many minor variations in which the bore diameter is varied and the yoke shape altered, but the optical behavior is not greatly affected. The design adopted is usually a compromise between the optical performance desired and the technological needs of the user. In high-performance systems, the specimen is usually inside the field region and may be inserted either down the upper bore (top entry) or laterally through the gap (side entry). The specimen-holder mechanism requires a certain volume, especially if it is of one of the sophisticated models that permit *in situ* experiments: specimen heating, to study phase changes in alloys, for example, or specimen cooling to liquid nitrogen or liquid helium temperature, or straining; specimen rotation and tilt are routine requirements of the metallurgist. All this requires space in the gap region, which is further encumbered by a cooling device to protect the specimen from contamination, the stigmator, and the objective aperture drive. The desired optical properties must be achieved subject to the demands on space of all these devices, as far as this is possible. As Ugo Valdrè has said, the interior of an electron microscope objective should be regarded as a microlaboratory.

Magnetic lenses commonly operate at room temperature, but there is some advantage in going to very

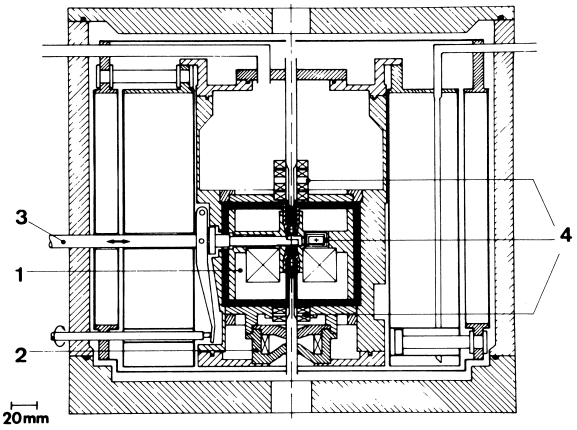


FIGURE 14 Superconducting lens system: (1) objective (shielding lens); (2) intermediate with iron circuit; (3) specimen holder; and (4) corrector device.

low temperature and running in the superconducting regime. Several designs have been explored since Andrè Laberrigue, Humberto Fernández-Morán, and Hans Boersch introduced the first superconducting lenses, but only one has survived, the superconducting shielding lens introduced by Isolde Dietrich and colleagues at Siemens (Fig. 14). Here, the entire lens is at a very low temperature, the axial field being produced by a superconducting coil and concentrated into the narrow gap region by superconducting tubes. Owing to the Meissner–Ochsenfeld effect, the magnetic field cannot penetrate the metal of these superconducting tubes and is hence concentrated in the gap. The field is likewise prevented from escaping from the gap by a superconducting shield. Such lenses have been incorporated into a number of microscopes and are particularly useful for studying material that must be examined at extremely low temperatures; organic specimens that are irretrievably damaged by the electron beam at higher temperatures are a striking example.

Despite their very different technology, these superconducting lenses have essentially the same optical properties as their warmer counterparts. This is not true of the various magnetic lenses that are grouped under the heading of unconventional designs; these were introduced mainly by Tom Mulvey, although the earliest, the minilens, was devised by Jan Le Poole. The common feature of these lenses, which are extremely varied in appearance, is that the space occupied by the lens is very different in volume or shape from that required by a traditional lens. A substantial reduction in the volume can be achieved by increasing the current density in the coil; in the minilens (Fig. 15), the value may be $\sim$80 mm^2, whereas in a conventional lens, 2 A/mm^2 is a typical figure. Such lenses are employed as auxiliary lenses in zones already occupied by other elements, such as bulky traditional lenses. After the initial success of these minilenses, a family of miniature lenses came into being, with which it would be possible to reduce the dimensions of the huge, heavy lenses used for very high voltage microscopes (in the megavolt range). Once the conventional design had been questioned, it was natural to inquire whether there was any advantage to be gained by abandoning its symmetric shape. This led to the invention of the pancake lens, flat like a phonograph record, and various single-polepiece or "snorkel" lenses (Fig. 16). These are attractive in situations where the electrons are at the end of their trajectory, and the single-polepiece design of Fig. 16 can be used with a target in front of it or a gun beyond it. Owing to their very flat shape, such lenses, with a bore, can be used to render microscope projector systems free of certain distortions, which are otherwise very difficult to eliminate.

This does not exhaust all the types of magnetic lens. For many years, permanent-magnet lenses were investigated in the hope that a simple and inexpensive microscope could be constructed with them. An addition to the family of traditional lenses is the unsymmetric triple-polepiece lens, which offers the same advantages as the single-polepiece designs in the projector system. Magnetic lens studies have also been revivified by the needs of electron beam lithography.

4. Aberration Correction

The quest for high resolution has been a persistent preoccupation of microscope designers since these instruments came into being. Scherzer's theorem (1936) was therefore a very unwelcome result, showing as it did that the principal resolution-limiting aberration could never vanish in

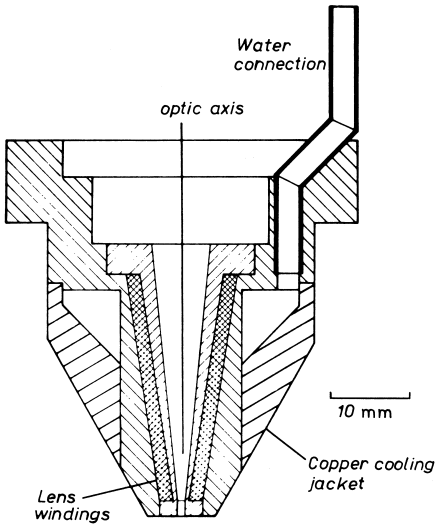


FIGURE 15 Minilens.

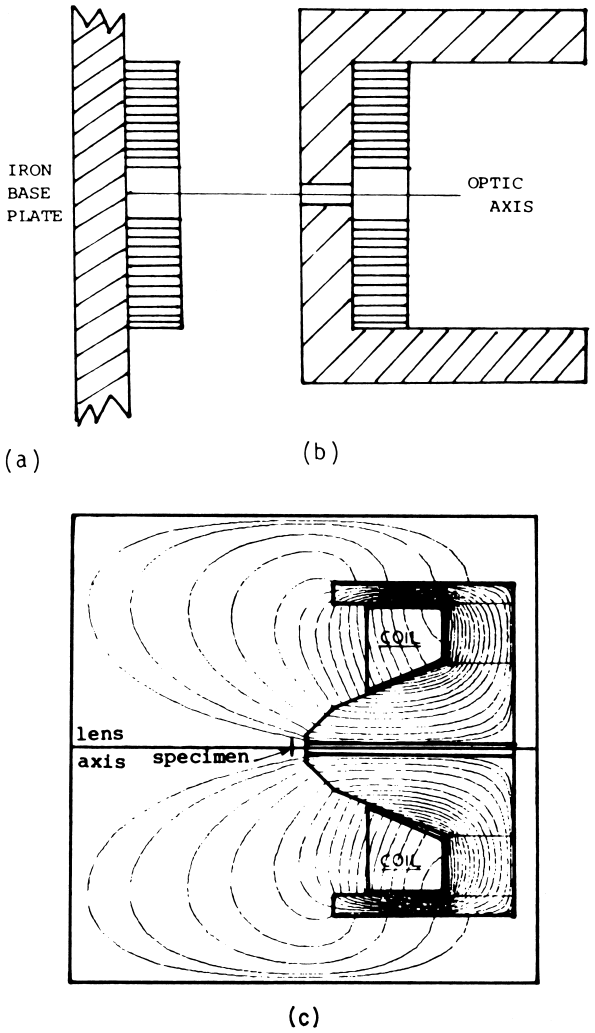


FIGURE 16 Some unconventional magnetic lenses.

round lenses. It was Scherzer again who pointed out (1947) the various ways of circumventing his earlier result by introducing aberration correctors of various kinds. The proof of the theorem required rotational symmetry, static fields, the absence of space charge, and the continuity of certain properties of the electrostatic potential. By relaxing any one of these requirements, aberration correction is in principle possible, but only two approaches have achieved any measure of success.

The most promising type of corrector was long believed to be that obtained by departing from rotational symmetry, and it was with such devices that correction was at last successfully achieved in the late 1990s. Such correctors fall into two classes. In the first, quadrupole lenses are employed. These introduce new aperture aberrations, but by adding octopole fields, the combined aberration of the round lens and the quadrupoles can be cancelled. At least four quadrupoles and three octopoles are required.

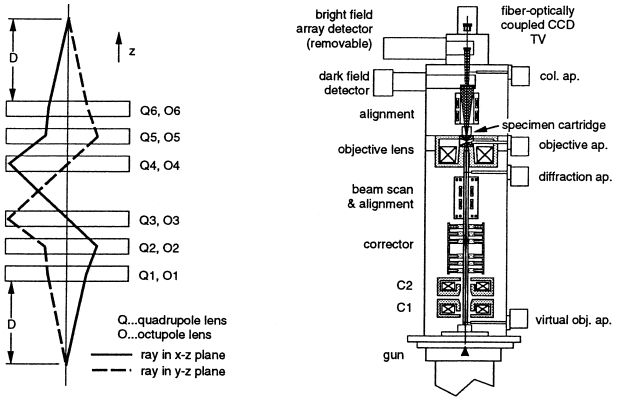


FIGURE 17 Correction of spherical aberration in a scanning transmission electron microscope. (Left) Schematic diagram of the quadrupole–octopole corrector and typical trajectories. (Right) Incorporation of the corrector in the column of a Vacuum Generators STEM. [From Krivanek, O. L., et al. (1997). Institute of Physics Conference Series 153, 35. Copyright IOP Publishing.]

A corrector based on this principle has been incorporated into a scanning transmission electron microscope by O. Krivanek at the University of Cambridge (Fig. 17). In the second class of corrector, the nonrotationally symmetric elements are sextupoles. A suitable combination of two sextupoles has a spherical aberration similar to that of a round lens but of opposite sign, and the undesirable second-order aberrations cancel out (Fig. 18). The technical difficulties of introducing such a corrector in a high-resolution transmission electron microscope have been overcome by M. Haider (Fig. 19).

Quadrupoles and octopoles had seemed the most likely type of corrector to succeed because the disturbance to the existing instrument, already capable of an unaided resolution of a few angstroms, was slight. The family of correctors that employ space charge or charged foils placed across the beam perturb the microscope rather more. Efforts continue to improve lenses by inserting one or more

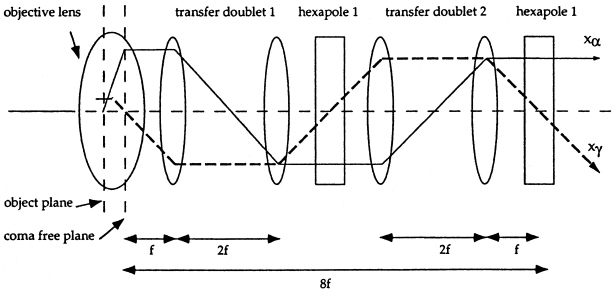


FIGURE 18 Correction of spherical aberration in a transmission electron microscope. Arrangemnent of round lenses and sextupoles (hexapoles) that forms a semiaplanatic objective lens. The distances are chosen to eliminate radial coma. [From Haider, M., et al. (1995). Optik 99, 167. Copyright Wissenschaftliche Verlagsgesellschaft.]

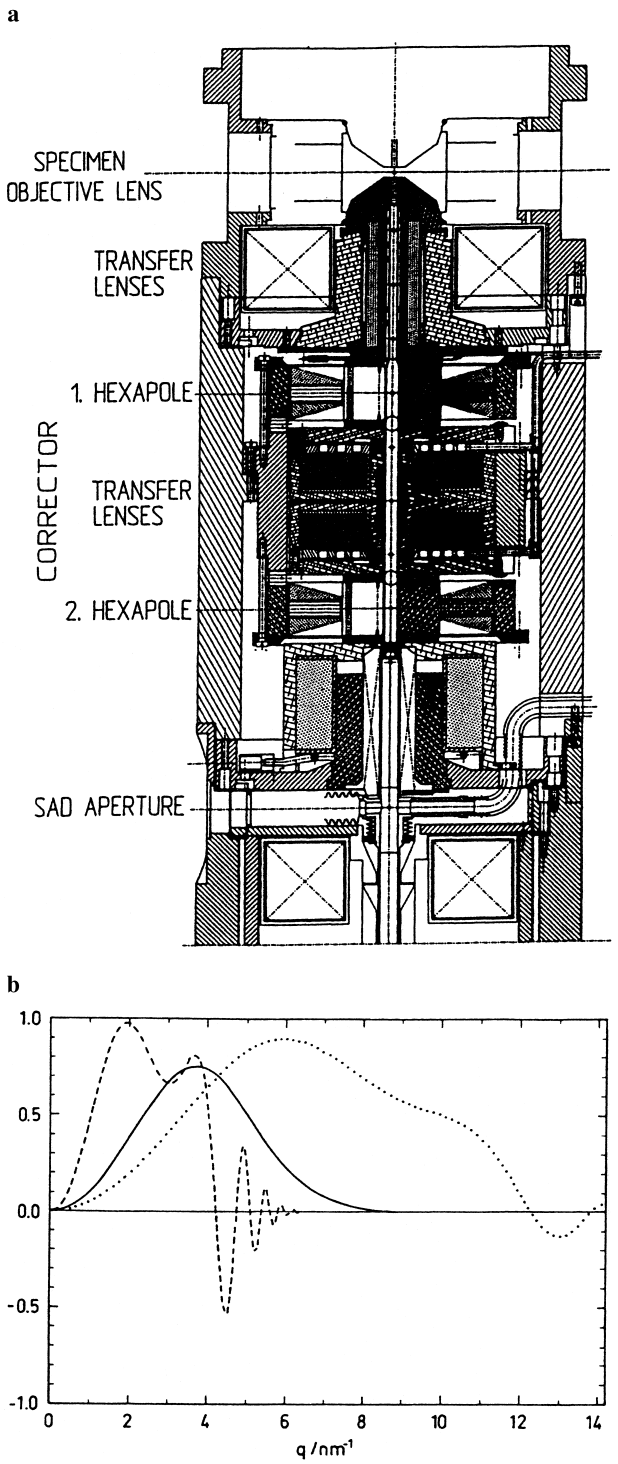


FIGURE 19 (a) The corrector of Fig. 18 incorporated in a transmission electron microscope. (b) The phase contrast transfer function of the corrected microscope. Dashed line: no correction. Full line: corrector switched on, energy width (a measure of the temporal coherence) 0.7 eV. Dotted line: energy width 0.2 eV. Chromatic aberration remains a problem, and the full benefit of the corrector is obtained only if the energy width is very narrow. [From Haider, M., et al. (1998). J. Electron Microsc. 47, 395. Copyright Japanese Society of Electron Microscopy.]

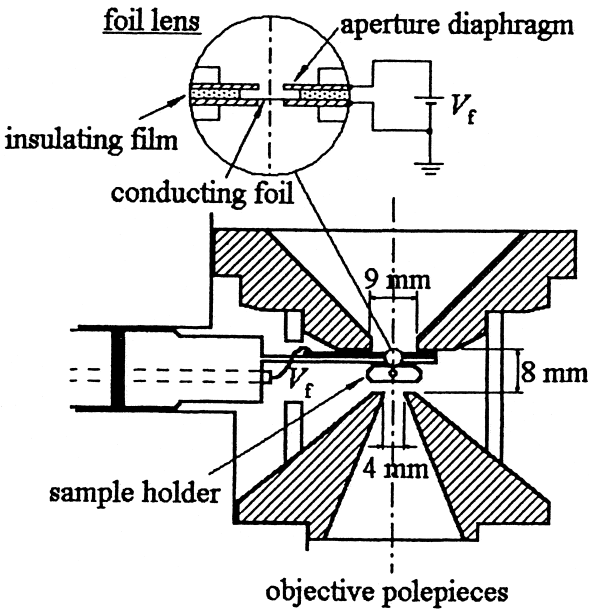


FIGURE 20 Foil lens and polepieces of an objective lens to be corrected. [From Hanai, T., et al. (1998). J. Electron Microsc. 47, 185. Copyright Japanese Society of Electron Microscopy.]

foils in the path of the electrons, with a certain measure of success, but doubts still persist about this method. Even if a reduction in total C_s is achieved, the foil must have a finite thickness and will inevitably scatter the electrons traversing it. How is this scattering to be separated from that due to the specimen? Figure 20 shows the design employed in an ongoing Japanese project.

An even more radical solution involves replacing the static objective lens by one or more microwave cavities. In Scherzer's original proposal, the incident electron beam was broken into short pulses and the electrons far from the axis would hence arrive at the lens slightly later than those traveling near the axis. By arranging that the axial electrons encounter the maximum field so that the peripheral electrons experience a weaker field, Scherzer argued, the effect of C_s could be eliminated since, in static lenses, the peripheral electrons are too strongly focused. Unfortunately, when we insert realistic figures into the corresponding equations, we find that the necessary frequency is in the gigahertz range, with the result that the electrons spend a substantial part of a cycle, or more than a cycle, within the microwave field. Although this means that the simple explanation is inadequate, it does not invalidate the principle, and experiment and theory both show that microwave cavity lenses can have positive or negative spherical aberration coefficients. The principal obstacles to their use are the need to produce very short pulses containing sufficient current and, above all, the fact that the beam emerging from such cavity lenses has a rather large energy

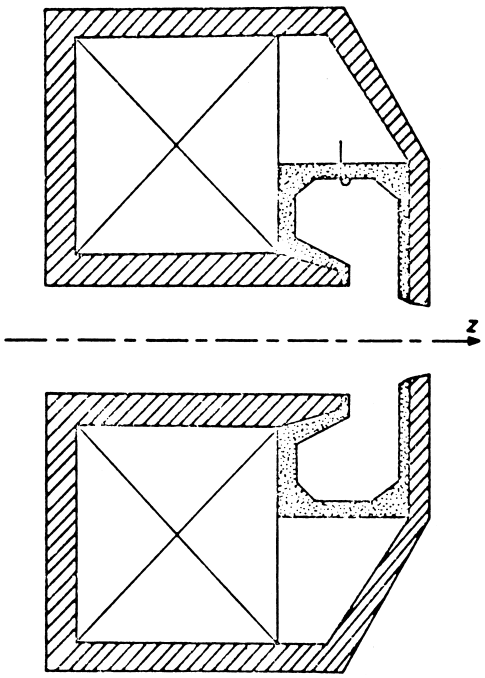


FIGURE 21 Microwave cavity lens between the polepieces of a magnetic lens. (Courtesy of L. C. Oldfield.)

spread, which makes further magnification a problem. An example is shown in Fig. 21.

Finally, we mention the possibility of *a posteriori* correction in which we accept the deleterious effect of C_s on the recorded micrograph but attempt to reduce or eliminate it by subsequent digital or analog processing of the image. A knowledge of the wave theory of electron image formation is needed to understand this idea and we therefore defer discussion of it to Section III.B.

5. Prisms, Mirrors, and Energy Analyzers

Magnetic and electrostatic prisms and systems built up from these are used mainly for their dispersive properties in particle optics. We have not yet encountered electron mirrors, but we mention them here because a mirror action is associated with some prisms; if electrons encounter a potential barrier that is high enough to halt them, they will be reflected and a paraxial optical formalism can be developed to describe such mirror optics. This is less straightforward than for lenses, since the ray gradient is far from small at the turning point, which means that one of the usual paraxial assumptions that off-axis distance and ray gradient are everywhere small is no longer justified.

The simplest magnetic prisms, as we have seen, are sector fields created by magnets of the C-type or picture-frame arrangement (Fig. 22) with circular poles or sector poles with a sector or rectangular yoke. These or analogous electrostatic designs can be combined in many ways, of which we can mention only a small selection. A very ingenious arrangement, which combines magnetic deflection with an electrostatic mirror, is the Castaing–Henry analyzer (Figs. 23a–23c) which has the constructional convenience that the incident and emergent optic axes are in line; its optical properties are such that an energy-filtered image or an energy spectrum from a selected area can be obtained. A natural extension of this is the magnetic Ω filter (Fig. 23d), in which the mirror is suppressed; if the particle energy is not too high, use of the electrostatic analog of this can be envisaged (Fig. 23e). It is possible to eliminate many of the aberrations of such filters by arranging the system not only symmetrically about the mid-plane ($x' - x$ in Fig. 23d), but also antisymmetrically about the planes midway between the mid-plane and the optic axis. A vast number of prism combinations have been explored by Veniamin Kel'man and colleagues in Alma-Ata in the quest for high-performance mass and electron spectrometers.

Energy analysis is a subject in itself, and we can do no more than mention various other kinds of energy or

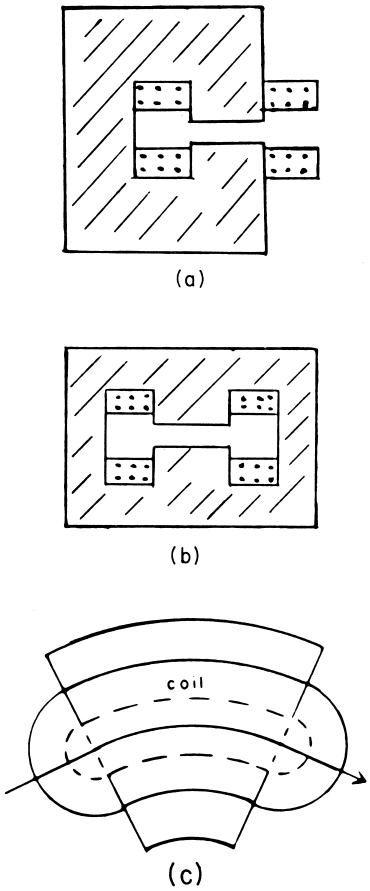


FIGURE 22 (a) C-Type and (b) picture-frame magnets URE typically having (c) sector-shaped yoke and poles.

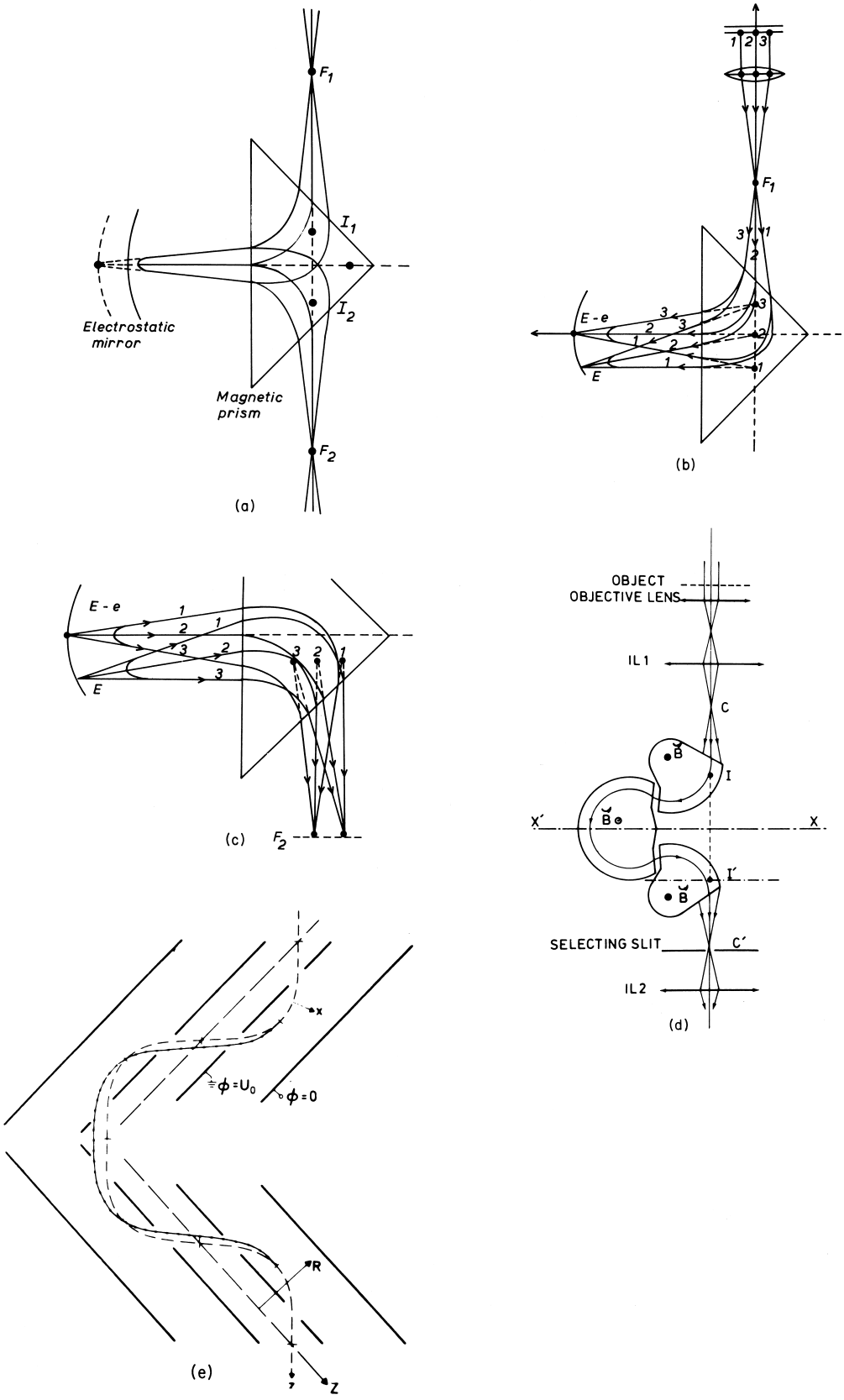


FIGURE 23 Analyzers: (a–c) Castaing–Henry analyzer; (d) Ω filter; and (e) electrostatic analog of the Ω filter.

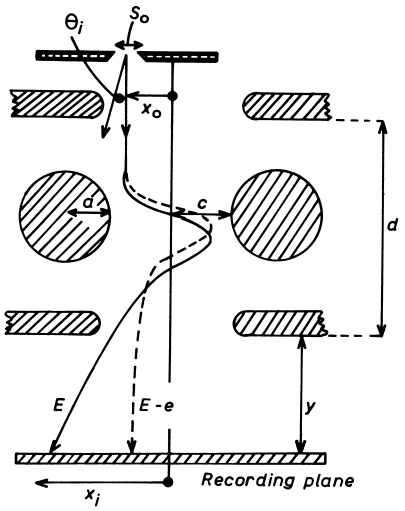


FIGURE 24 Möllenstedt analyzer.

momentum analyzers. The Wien filter consists of crossed electrostatic and magnetic fields, through which particles of a particular energy will pass undeflected, whereas all others will be deviated from their path. The early β-ray spectrometers exploited the fact that the chromatic aberration of a lens causes particles of different energies to be focused in different planes. The Möllenstedt analyzer is based on the fact that rays in an electrostatic lens far from the axis are rapidly separated if their energies are different (Fig. 24). The Ichinokawa analyzer is the magnetic analog of this and is used at higher accelerating voltages where electrostatic lenses are no longer practicable. In retarding-field analyzers, a potential barrier is placed in the path of the electrons and the current recorded as the barrier is progressively lowered.

6. Combined Deflection and Focusing Devices

In the quest for microminiaturization, electron beam lithography has acquired considerable importance. It proves to be advantageous to include focusing and deflecting fields within the same volume, and the optical properties of such combined devices have hence been thoroughly studied, particularly, their aberrations. It is important to keep the adverse effect of these aberrations small, especially because the beam must be deflected far from the original optical axis. An ingenious way of achieving this, proposed by Hajime, Ohiwa, is to arrange that the optic axis effectively shifts parallel to itself as the deflecting field is applied; for this, appropriate additional deflection, round and multipole fields must be superimposed and the result may be regarded as a "moving objective lens" (MOL) or "variable-axis lens" (VAL). Perfected immersion versions of these and of the "swinging objective lens" (SOL) have been developed, in which the target lies within the field region.

III. WAVE OPTICS

A. Wave Propagation

The starting point here is not the Newton–Lorentz equations but Schrödinger's equation; we shall use the nonrelativistic form, which can be trivially extended to include relativistic effects for magnetic lenses. Spin is thus neglected, which is entirely justifiable in the vast majority of practical situations. The full Schrödinger equation takes the form

$$-\frac{\hbar^2}{m_0}\nabla^2\Psi + \frac{e\hbar}{im_0}\mathbf{A}\cdot\operatorname{grad}\Psi + \left(-e\Phi + \frac{e^2}{2m_0}\mathbf{A}^2\right)\Psi - i\hbar\frac{\partial\Psi}{\partial t} = 0 \tag{54}$$

and writing

$$\Psi(x, y, z, t) = \psi(x, y, z)e^{-i\omega t} \tag{55}$$

we obtain

$$-\frac{\hbar^2}{2m_0}\nabla^2\psi + \frac{e\hbar}{im_0}\mathbf{A}\cdot\operatorname{grad}\psi + \left(-e\Phi + \frac{e^2}{2m_0}\mathbf{A}^2\right)\psi = E\psi \tag{56}$$

with

$$E = \hbar\omega \tag{57}$$

where $\hbar = h/2\pi$ and h is Planck's constant. The free-space solution corresponds to

$$p = h/\lambda$$

or

$$\lambda = h/(2em_0\phi_0)^{1/2} \approx 12.5/\phi_0^{1/2} \tag{58}$$

where p is the momentum.

As in the case of geometric optics, we consider the paraxial approximation, which for the Schrödinger equation takes the form

$$-\hbar^2\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}\right) + \frac{1}{2}em_0(\phi'' + \eta^2B^2)(x^2 + y^2)\psi - i\hbar p'\psi - 2i\hbar p\frac{\partial\psi}{\partial z} = 0 \tag{59}$$

and we seek a wavelike solution:

$$\psi(x, y, z) = a(z)\exp[iS(x, y, z)/\hbar]. \tag{60}$$

After some calculation, we obtain the required equation describing the propagation of the wave function through electrostatic and magnetic fields:

$$\psi(x, y, z) = \frac{p_0^{3/2}}{2\pi i\hbar h(z)p^{1/2}} \exp\left[\frac{ipg'(z)}{2\hbar g(z)}(x^2 + y^2)\right] \times \iint_{-\infty}^{\infty} \psi(x_o, y_o, z_o) \exp\left\{\frac{ip_o}{2\hbar g(z)h(z)} \times \left[(x - x_o g)^2 + (y - y_o g)^2\right] dx_o\, dy_o \right. \tag{61}$$

This extremely important equation is the basis for all that follows. In it, $g(z)$ and $h(z)$ now denote the solutions of the geometric paraxial equations satisfying the boundary conditions $g(z_o) = h'(z_o) = 1$, $g'(z_o) = h(z_o) = 0$. Reorganizing the various terms, Eq. (61) can be written

$$\psi(x, y, z) = \frac{1}{i\lambda r h(z)} \iint_{-\infty}^{\infty} \psi(x_o, y_o, z_o) \times \exp\left\{\frac{i\pi}{\lambda h(z)}\left[g(z)\left(x_o^2 + y_o^2\right) - 2(x_o x + y_o y) + rh'(z)(x^2 + y^2)\right]\right\} dx_o\, dy_o \tag{62}$$

with $\lambda = h/p_o$ and $r = p/p_o = (\phi/\phi_o)^{1/2}$.

Let us consider the plane $z = z_d$ in which $g(z)$ vanishes, $g(z_d) = 0$. For the magnetic case ($r = 1$), we find

$$\psi(x_d, y_d, z_d) = \frac{E_d}{i\lambda h(z_o)} \iint \psi(x_o, y_o, z_o) \times \exp\left[-\frac{2i}{\lambda h(z_d)}(x_o x_d + y_o y_d)\right] dx_o\, dy_o \tag{63}$$

with $E_d = \exp[i\pi h'(z_d)(x_d^2 + y_d^2)/\lambda h(z_d)]$, so that, scale factors apart, the wave function in this plane is the Fourier transform of the same function in the object plane.

We now consider the relation between the wave function in the object plane and in the image plane $z = z_i$ conjugate to this, in which $h(z)$ vanishes: $h(z_i) = 0$. It is convenient to calculate this in two stages, first relating $\psi(x_i, y_i, z_i)$ to the wave function in the exit pupil plane of the lens, $\psi(x_a, y_a, z_a)$ and then calculating the latter with the aid of Eq. (62). Introducing the paraxial solutions $G(z)$, $H(z)$ such that

$$G(z_a) = H'(z_a) = 1; \qquad G'(z_a) = H(z_a) = 0$$

we have

$$\psi(x_i, y_i, z_i) = \frac{1}{i\lambda H(z_i)} \iint \psi(x_a, y_a, z_a) \times \exp\left\{\frac{i\pi}{\lambda H(z)}\left[G(z_i)\left(x_a^2 + y_a^2\right) - 2(x_a x_i + y_a y_i) + H'(z_i)\left(x_i^2 + y_i^2\right)\right]\right\} dx_a\, dy_a \tag{64}$$

Using Eq. (62), we find

$$M\psi(x_i, y_i, z_i)E_i = \iint \psi(x_o, y_o, z_o)K(x_i, y_i; x_o, y_o)E_o\, dx_o\, dy_o \tag{65}$$

where M is the magnification, $M = g(z_i)$, and

$$E_i = \exp\left[\frac{i\pi}{\lambda M}\frac{g_a h_i' - g_i' h_a}{h_a}\left(x_i^2 + y_i^2\right)\right]$$
$$E_o = \exp\left[\frac{i\pi g_a}{\lambda h_a}\left(x_o^2 + y_o^2\right)\right] \tag{66}$$

These quadratic factors are of little practical consequence; they measure the curvature of the wave surface arriving at the specimen and at the image. If the diffraction pattern plane coincides with the exit pupil, then $E_o = 1$. We write $h(z_a) = f$ since this quantity is in practice close to the focal length, so that for the case $z_d = z_a$,

$$E_i = \exp\left[-\frac{i\pi g_i'}{\lambda M}\left(x_i^2 + y_i^2\right)\right] \tag{67}$$

The most important quantity in Eq. (65) is the function $K(x_i, y_i; x_o, y_o)$, which is given by

$$K(x, y; x_o, y_o) = \frac{1}{\lambda^2 f^2} \iint A(x_a, y_a) \times \exp\left\{-\frac{2\pi i}{\lambda f}\left[\left(x_o - \frac{x}{M}\right)x_a + \left(y_o - \frac{y}{M}\right)y_a\right]\right\} dx_a\, dy_a \tag{68}$$

or introducing the spatial frequency components

$$\xi = x_a/\lambda f; \qquad \eta = y_a/\lambda f \tag{69}$$

we find

$$K(x, y; x_o, y_o) = \iint A(\lambda f\xi, \lambda f\eta) \times \exp\left\{-2\pi i\left[\left(x_o - \frac{x}{M}\right)\xi + \left(y_o - \frac{y}{M}\right)\eta\right]\right\} d\xi\, d\eta \tag{70}$$

In the paraxial approximation, the aperture function A is simply a mathematical device defining the area of integration in the aperture plane: $A = 1$ inside the pupil and $A = 0$ outside the pupil. If we wish to include the effect of geometric aberrations, however, we can represent them as a phase shift of the electron wave function at the exit pupil. Thus, if the lens suffers from spherical aberration, we write

$$A(x_a, y_a) = a(x_a, y_a) \exp[-i\gamma(x_a, y_a)] \tag{71}$$

in which

$$\gamma = \frac{2\pi}{\lambda}\left\{\frac{1}{4}C_s\left(\frac{x_a^2 + y_a^2}{f^2}\right)^2 - \frac{1}{2}\Delta\frac{x_a^2 + y_a^2}{f^2}\right\}$$
$$= \frac{\pi\lambda}{2}\left\{C_s\lambda^2(\xi^2 + \eta^2)^2 - 2\Delta(\xi^2 + \eta^2)\right\} \tag{72}$$

the last term in Δ allowing for any defocus, that is, any small difference between the object plane and the plane conjugate to the image plane. All the third-order geometric aberrations can be included in the phase shift γ, but we consider only C_s and the defocus Δ. This limitation is justified by the fact that C_s is the dominant aberration of objective lenses and proves to be extremely convenient because Eq. (65) relating the image and object wave functions then has the form of a convolution, which it loses if other aberrations are retained (although coma can be accommodated rather uncomfortably). It is now the amplitude function $a(x_a, y_a)$ that represents the physical pupil, being equal to unity inside the opening and zero elsewhere.

In the light of all this, we rewrite Eq. (65) as

$$E_i\psi(x_i, y_i, z_i) = \frac{1}{M}\iint K\left(\frac{x_i}{M} - x_o, \frac{y_i}{M} - y_o\right)E_o \times \psi_o(x_o, y_o, z_o)\,dx_o\,dy_o \tag{73}$$

Defining the Fourier transforms of ψ_o, ψ_i, and K as follows,

$$\begin{aligned}
\tilde{\psi}_o(\xi, \eta) &= \iint E_o\psi_o \times \exp[-2\pi i(\xi x_o + \eta y_o)]\,dx_o\,dy_o \\
\tilde{\psi}_o(\xi, \eta) &= \iint E_i\psi_i(Mx_i, My_i) \times \exp[-2\pi i(\xi x_i + \eta y_i)]\,dx_i\,dy_i \\
&= \frac{1}{M^2}\iint E_i\psi_i(x_i, y_i) \times \exp\left[-2\pi i\frac{(\xi x_i + \eta y_i)}{M}\right]dx_i\,dy_i \\
\tilde{K}(\xi, \eta) &= \iint K(x, y) \times \exp[-2\pi i(\xi x + \eta y)]\,dx\,dy
\end{aligned} \tag{74}$$

in which small departures from the conventional definitions have been introduced to assimilate inconvenient factors, Eq. (65) becomes

$$\tilde{\psi}_i(\xi, \eta) = \frac{1}{M}\tilde{K}(\xi, \eta)\tilde{\psi}_o(\xi, \eta) \tag{75}$$

This relation is central to the comprehension of electron-optical image-forming instruments, for it tells us that the formation of an image may be regarded as a filtering operation. If $\tilde{K}$ were equal to unity, the image wave function would be identical with the object wave function, appropriately magnified; but in reality $\tilde{K}$ is not unity and different spatial frequencies of the wave leaving the specimen, $\psi(x_o, y_o, z_o)$, are transferred to the image with different weights. Some may be suppressed, some attenuated, some may have their sign reversed, and some, fortunately, pass through the filter unaffected. The notion of spatial frequency is the spatial analog of the temporal frequency, and we associate high spatial frequencies with fine detail and low frequencies with coarse detail; the exact interpretation is in terms of the fourier transform, as we have seen.

We shall use Eqs. (73) and (75) to study image formation in two types of optical instruments, the transmission electron microscope (TEM) and its scanning transmission counterpart, the STEM. This is the subject of the next section.

B. Instrumental Optics: Microscopes

The conventional electron microscope (TEM) consists of a source, condenser lenses to illuminate a limited area of the specimen, an objective to provide the first stage of magnification beyond the specimen, and projector lenses, which magnify the first intermediate image or, in diffraction conditions, the pattern formed in the plane denoted by $z = z_d$ in the preceding section. In the STEM, the role of the condenser lenses is to demagnify the crossover so that a very small electron probe is formed on the specimen. Scanning coils move this probe over the surface of the latter in a regular raster, and detectors downstream measure the current transmitted. There are inevitably several detectors, because transmission microscope specimens are essentially transparent to electrons, and thus there is no diminution of the total current but there is a redistribution of the directions of motion present in the beam. Electron-optical specimens deflect electrons but do not absorb them. In the language of light optics, they are phase specimens, and the electron microscope possesses means of converting invisible phase variations ot amplitude variations that the eye can see.

We now examine image formation in the TEM in more detail. We first assume, and it is a very reasonable first approximation, that the specimen is illuminated by a parallel

uniform beam of electrons or, in other words, that the wave incident on the specimen is a constant. We represent the effect of the specimen on this wave by a multiplicative specimen transparency function $S(x_o, y_o)$, which is a satisfactory model for the very thin specimens employed for high-resolution work and for many other specimens. This specimen transparency is a complex function, and we write

$$S(x_o, y_o) = [1 - s(x_o, y_o)] \exp[i\varphi(x_o, y_o)] \quad (76a)$$

$$= 1 - s + i\varphi \quad (76b)$$

for small values of s and φ. The real term s requires some explanation, for our earlier remarks suggest that s must vanish if no electrons are halted by the specimen. we retain the term in s for two reasons. First, some electrons may be scattered inelastically in the specimen, in which case they must be regarded as lost in this simple monochromatic and hence monoenergetic version of the theory. Second, all but linear terms have been neglected in the approximate expression (76b) and, if necessary, the next-higher-order term $(-\frac{1}{2}\varphi^2)$ can be represented by s.

The wave leaving the specimen is now proportional to S normalizing, so that the constant of proportionality is unity; after we substitute

$$\psi(x_o, y_o, z_o) = 1 - s + i\varphi \quad (77)$$

into Eq. (75). Again denoting Fourier transforms by the tilde, we have

$$\tilde{\psi}_i(\xi, \eta) = \frac{1}{M}\tilde{K}(\xi, \eta)[\delta(\xi, \eta) - \tilde{s}(\xi, \eta) + i\tilde{\varphi}(\xi, \eta)]$$

$$= \frac{1}{M} a \exp(-i\gamma)(\delta - \tilde{s} + i\tilde{\varphi}) \quad (78)$$

and hence

$$\psi_i(Mx_i, My_i) = \frac{1}{M}\iint a \exp(-i\gamma)(\delta - \tilde{s} + i\tilde{\varphi}) \times \exp[2\pi i(\xi x_i + \eta y_i)]\, d\xi\, d\eta \quad (79)$$

The current density at the image, which is what we see on the fluorescent screen of the microscope and record on film, is proportional to $\psi_i\psi_i^*$. We find that if both φ and s are small,

$$M^2\psi_i\psi_i^* \approx 1 - 2\int_{-\infty}^{\infty}\!\!\iint a\tilde{s}\cos\gamma \times \exp[2\pi i(\xi x + \eta y)]\, d\xi\, d\eta + 2\int_{-\infty}^{\infty}\!\!\iint a\tilde{\varphi}\sin\gamma \times \exp[2\pi i(\xi x + \eta y)]\, d\xi\, d\eta \quad (80)$$

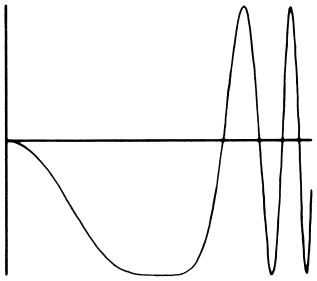


FIGURE 25 Function $\sin\gamma$ at Scherzer defocus $\Delta = (C_s\lambda)^{1/2}$.

and writing $j = M^2\psi_i\psi_i^*$ and $C = j - 1$, we see that

$$\tilde{C} = -2a\tilde{s}\cos\gamma + 2a\tilde{\varphi}\sin\gamma \quad (81)$$

This justifies our earlier qualitative description of image formation as a filter process. Here we see that the two families of spatial frequencies characterizing the specimen, $\tilde{\varphi}$ and $\tilde{s}$, are distorted before they reach the image by the linear filters $\cos\gamma$ and $\sin\gamma$. The latter is by far the more important. A typical example is shown in Fig. 25. The distribution $2\sin\gamma$ can be observed directly by examining an amorphous phase specimen, a very thin carbon film, for example. The spatial frequency spectrum of such a specimen is fairly uniform over a wide range of frequencies so that $\tilde{C} \infty \sin\gamma$. A typical spectrum is shown in Fig. 26, in which the radial intensity distribution is proportional to $\sin^2\gamma$. Such spectra can be used to estimate the defocus Δ and the coefficient C_s very accurately.

The foregoing discussion is idealized in two respects, both serious in practice. First, the illuminating beam has been assumed to be perfectly monochromatic, whereas in reality there will be a spread of wavelengths of several parts per million; in addition, the wave incident on the specimen has been regarded as a uniform plane wave, which is equivalent to saying that it originated in an ideal ponint source. Real sources, of course, have a finite size, and the single plane wave should therefore be replaced by a spectrum of plane waves incident at a range of small angles to the specimen. The step from point source and monochromatic beam to finite source size and finite wavelength spread is equivalent to replacing perfectly coherent illumination by partially coherent radiation, with the wavelength spread corresponding to temporal partial coherence and the finite source size corresponding to spatial partial coherence. (We cannot discuss the legitimacy of separating these effects here, but simply state that this is almost always permissible.) Each can be represented by an envelope function, which multiplies the coherent transfer functions $\sin\gamma$ and $\cos\gamma$. This is easily seen for the temporal spatial coherence. Let us associate a probability distribution $H(f)$, $\int H(f)\, df = 1$, with the current density at each point in the image, the argument f being some

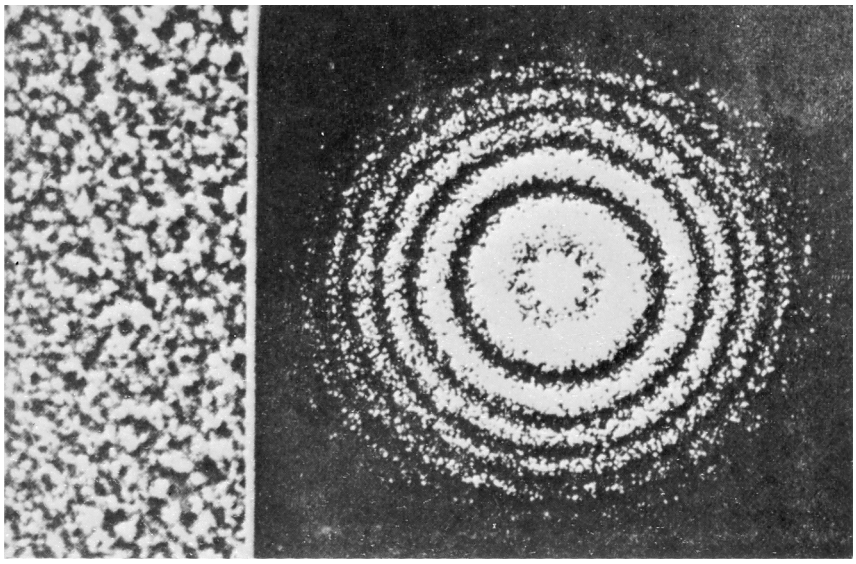


FIGURE 26 Spatial frequency spectrum (right) of an amorphous phase specimen (left).

convenient measure of the energy variation in the beam incident on the specimen. Hence, $dj/j = H(f)\,df$. From Eq. (80), we find

$$j = 1 - \int a\tilde{s} T_s \exp[2\pi i(\xi x + \eta y)]\,d\xi\,d\eta + \int a\tilde{\varphi} T_\varphi \exp[2\pi i(\xi x + \eta y)]\,d\xi\,d\eta \qquad (82)$$

where

$$T_s = 2\int \cos\gamma(\xi, \eta, f)H(f)\,df$$
$$T_\varphi = 2\int \sin\gamma(\xi, \eta, f)H(f)\,df \qquad (83)$$

and if f is a measure of the defocus variation associated with the energy spread, we may set Δ equal to $\Delta_o + f$, giving

$$T_s = 2\cos\gamma \int H(f)\cos[\pi\lambda f(\xi^2 + \eta^2)]\,df$$
$$T_\varphi = 2\sin\gamma \int H(f)\cos[\pi\lambda f(\xi^2 + \eta^2)]\,df \qquad (84)$$

if $H(f)$ is even, and a slightly longer expression when it is not.

The familiar $\sin\gamma$ and $\cos\gamma$ are thus clearly seen to be modulated by an envelope function, which is essentially the Fourier transform of $H(f)$. A similar result can be obtained for the effect of spatial partial coherence, but the demonstration is longer. Some typical envelope functions are shown in Fig. 27.

An important feature of the function $\sin\gamma$ is that it gives us a convenient measure of the resolution of the microscope. Beyond the first zero of the function, information is no longer transferred faithfully, but in the first zone the transfer is reasonably correct until the curve begins to dip toward zero for certain privileged values of the defocus, $\Delta = (C_s\lambda)^{1/2}$, $(3C_s\lambda)^{1/2}$, and $(5C_s\lambda)^{1/2}$; for the first of these values, known as the Scherzer defocus,

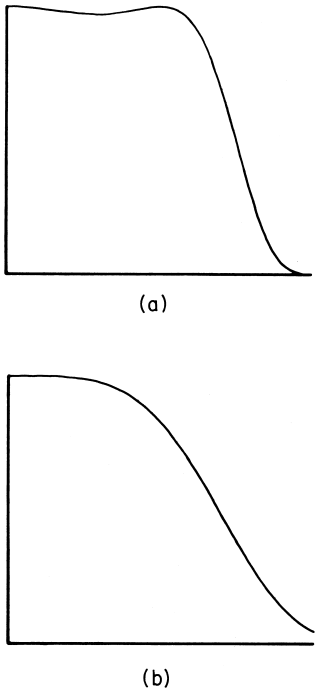


FIGURE 27 Envelope functions characterizing (a) spatial and (b) temporal partial coherence.

the zero occurs at the spatial frequency $(C_s\lambda^3)^{-1/4}$; the reciprocal of this multiplied by one of various factors has long been regarded as the resolution limit of the electron microscope, but transfer function theory enables us to understand the content of the image in the vicinity of the limit in much greater detail. The arrival of commercial electron microscopes equipped with spherical aberration correctors is having a profound influence on the practical exploitation of transfer theory. Hitherto, the effect of spherical aberration dictated the mode of operation of the TEM when the highest resolution was required. When the coefficient of spherical aberration has been rendered very small by correction, this defect is no longer the limiting factor and other modes of operation become of interest.

We now turn to the STEM. Here a bright source, typically a field-emission gun, is focused onto the specimen; the small probe is scanned over the surface and, well beyond the specimen, a far-field diffraction pattern of each elementary object area is formed. This pattern is sampled by a structured detector, which in the simplest case consists of a plate with a hole in the center, behind which is another plate or, more commonly, an energy analyzer. The signals from the various detectors are displayed on cathode-ray tubes, locked in synchronism with the scanning coils of the microscope. The reason for this combination of annular detector and central detector is to be found in the laws describing electron scattering. The electrons incident on a thin specimen may pass through unaffected; or they may be deflected with virtually no transfer of energy to the specimen, in which case they are said to be elastically scattered; or they may be deflected and lose energy, in which case they are inelastically scattered. The important point is that, on average, inelastically scattered electrons are deflected through smaller angles than those scattered elastically, with the result that the annular detector receives mostly elastically scattered particles, whereas the central detector collects those that have suffered inelastic collisions. The latter therefore have a range of energies, which can be separated by means of an energy analyzer, and we could, for example, form an image with the electrons corresponding to the most probable energy loss for some particular chemical element of interest. Another imaging mode exploits electrons that have been Rutherford scattered through rather large angles.

These modes of STEM image formation and others that we shall meet below can be explained in terms of a transfer function theory analogous to that derived for the TEM. This is not surprising, for many of the properties of the STEM can be understood by regarding it as an inverted TEM, the TEM gun corresponding to the small central detector of the STEM and the large recording area of the TEM to the source in the STEM, spread out over a large zone if we project back the scanning probe. We will not pursue this analogy here, but most texts on the STEM explore it in some detail. Consider now a probe centered on a point $\mathbf{x}_o = \boldsymbol{\xi}$ in the specimen plane of the STEM. We shall use a vector notation here, so that $\mathbf{x}_o = (x_o, y_o)$, and similarly for other coordinates. The wave emerging from the specimen will be given by

$$\psi(\mathbf{x}_o; \boldsymbol{\xi}) = S(\mathbf{x}_o)K(\boldsymbol{\xi} - \mathbf{x}_o) \tag{85}$$

in which $S(\mathbf{x}_o)$ is again the specimen transparency and K describes the incident wave and, in particular, the effect of the pupil size, defocus, and aberrations of the probe-forming lens, the last member of the condenser system. Far below the specimen, in the detector plane (subscript d) the wave function is given by

$$\psi_d(\mathbf{x}_d, \boldsymbol{\xi}) = \int S(\mathbf{x}_o)K(\boldsymbol{\xi} - \mathbf{x}_o) \times \exp(-2\pi i \mathbf{x}_d \cdot \mathbf{x}_o/\lambda R)\, d\mathbf{x}_o \tag{86}$$

in which R is a measure of the effective distance between the specimen and the detector. The shape of the detector (and its response if this is not uniform) can most easily be expressed by introducing a detector function $D(\mathbf{x}_d)$, equal to zero outside the detector and equal to its response, usually uniform, over its surface. The detector records incident current, and the signal generated is therefore proportional to

$$\begin{aligned} j_d(\boldsymbol{\xi}) &= \int |\psi_d(\mathbf{x}_d; \boldsymbol{\xi})|^2 D(\mathbf{x}_d)\, d\mathbf{x}_d \\ &= \iiint S(\mathbf{x}_o)S^*(\mathbf{x}'_o)K(\boldsymbol{\xi} - \mathbf{x}_o)K^*(\boldsymbol{\xi} - \mathbf{x}'_o) \\ &\quad \times \exp[-2\pi i \mathbf{x}_d \cdot (\mathbf{x}_o - \boldsymbol{x}'_o)/\lambda R] \\ &\quad \times D(\mathbf{x}_d)\, d\mathbf{x}_o\, d\mathbf{x}'_o\, d\mathbf{x}_d \end{aligned} \tag{87}$$

or introducing the Fourier transform of the detector response,

$$\begin{aligned} j_d(\boldsymbol{\xi}) &= \iint S(\mathbf{x}_o)S^*\mathbf{x}'_o)K(\boldsymbol{\xi} - \mathbf{x}_o)K^*(\boldsymbol{\xi} - \mathbf{x}'_o) \\ &\quad \times \tilde{D}\left(\frac{\mathbf{x}_o - \mathbf{x}'_o}{\lambda R}\right) d\mathbf{x}_o\, d\mathbf{x}'_o \end{aligned} \tag{88}$$

We shall use the formula below to analyze the signals collected by the simpler detectors, but first we derive the STEM analog of the filter Eq. (81). For this we introduce

the Fourier transforms of S and K into the expression for $\psi_d(\mathbf{x}_d, \boldsymbol{\xi})$. Setting $\mathbf{u} = \mathbf{x}_d/\lambda R$, we obtain

$$\begin{aligned}\psi_d(\lambda R\mathbf{u};\boldsymbol{\xi}) &= \int S(\mathbf{x}_o)K(\boldsymbol{\xi}-\mathbf{x}_o)\exp(-2\pi i\mathbf{u}\cdot\mathbf{x}_o)\,d\mathbf{x}_o\\ &= \iiint \tilde{S}(\mathbf{p})\tilde{K}(\mathbf{q})\\ &\quad\times\exp[-2\pi i\mathbf{x}_o\cdot(\mathbf{u}-\mathbf{p}+\mathbf{q})]\\ &\quad\times\exp[(-2\pi i\mathbf{q}\cdot\boldsymbol{\xi})\,d\mathbf{p}\,d\mathbf{q}\,d\mathbf{x}_o\\ &= \iint \tilde{S}(\mathbf{p})\tilde{K}(\mathbf{q})\delta(\mathbf{u}-\mathbf{p}+\mathbf{q})\\ &\quad\times\exp(2\pi i\mathbf{q}\cdot\boldsymbol{\xi})\,d\mathbf{p}\,d\mathbf{q}\\ &= \int \tilde{S}(\mathbf{p})\tilde{K}(\mathbf{p}-\mathbf{u})\exp[2\pi i\boldsymbol{\xi}\cdot(\mathbf{p}-\mathbf{u})]\,d\mathbf{p}\end{aligned} \tag{89}$$

After some calculation, we obtain an expression for $j_d(\boldsymbol{\xi}) = \int j(\mathbf{x}_d;\boldsymbol{\xi})D(\mathbf{x}_d)\,d\mathbf{x}_d$ and hence for its Fourier transform

$$\tilde{j}_d(\mathbf{p}) = \int \tilde{j}(\mathbf{x}_d;\mathbf{p})D(\mathbf{x}_d)\,d\mathbf{x}_d \tag{90}$$

Explicitly,

$$\begin{aligned}\tilde{j}_d(\mathbf{p}) &= \int |\tilde{K}(\mathbf{x}_d/\lambda R)|^2 D(\mathbf{x}_d)\delta(\mathbf{p})\\ &\quad - \tilde{s}(\mathbf{p})\int q_s(\mathbf{x}_d/\lambda R;\mathbf{p}\,D(\mathbf{x}_d)\,d\mathbf{x}_d\\ &\quad + i\tilde{\varphi}(\mathbf{p})\int q_\varphi(\mathbf{x}_d/\lambda R;\mathbf{p})D(\mathbf{x}_d)\,d\mathbf{x}_d\end{aligned} \tag{91}$$

for weakly scattering objects, $s \ll 1$, $\varphi \ll 1$. The spatial frequency spectrum of the bright-field image signal is thus related to s and φ by a filter relation very similar to that obtained for the TEM.

We now return to Eqs. (87) and (88) to analyze the annular and central detector configuration. For a small axial detector, we see immediately that

$$j_d(\boldsymbol{\xi}) \propto \left| \int \tilde{S}(\mathbf{x}_o)K(\boldsymbol{\xi}-\mathbf{x}_o)\,d\mathbf{x}_o \right|^2 \tag{92}$$

which is very similar to the image observed in a TEM. For an annular detector, we divide $S(\mathbf{x}_o)$ into an unscattered and a scattered part, $S(\mathbf{x}_o) = 1 + \sigma_s(\mathbf{x}_o)$. The signal consists of two main contributions, one of the form $\int[\sigma_s(\mathbf{x}_o) + \sigma_s^*(\mathbf{x}_o))]\ |K(\boldsymbol{\xi}-\mathbf{x}_o)|^2\,d\mathbf{x}_o$, and the other $\int|\sigma_s(\mathbf{x}_o)|^2\ |K(\boldsymbol{\xi}-\mathbf{x}_o)|^2 d\mathbf{x}_o$. The latter term usually dominates.

We have seen that the current distribution in the detector plane at any instant is the far-field diffraction pattern of the object element currently illuminated. The fact that we have direct access to this wealth of information about the specimen is one of the remarkable and attractive features of the STEM, rendering possible imaging modes that present insuperable difficulties in the TEM. The simple detectors so far described hardly exploit this wealth of information at all, since only two total currents are measured, one falling on the central region, the other on the annular detector. A slightly more complicated geometry permits us to extract directly information about the phase variation $\varphi(\mathbf{x}_o)$ of the specimen transparency $S(\mathbf{x}_o)$. Here the detector is divided into four quadrants, and by forming appropriate linear combinations of the four signals thus generated, the gradient of the phase variation can be displayed immediately. This technique has been used to study the magnetic fields across domain boundaries in magnetic materials.

Other detector geometries have been proposed, and it is of interest that it is not necessary to equip the microscope with a host of different detectors, provided that the instrument has been interfaced to a computer. It is one of the conveniences of all scanning systems that the signal that serves to generate the image is produced sequentially and can therefore be dispatched directly to computer memory for subsequent or on-line processing if required. By forming the far-field diffraction pattern not on a single large detector but on a honeycomb of very small detectors and reading the currents detected by each cell into framestore memory, complete information about each elementary object area can be recorded. Framestore memory can be programmed to perform simple arithmetic operations, and the framestore can thus be instructed to multiply the incoming intensity data by 1 or 0 in such a way as to mimic any desired detector geometry. The signals from connected regions of the detector—quadrants, for example—are then added, and the total signal on each part is then stored, after which the operation is repeated for the next object element under the probe. Alternatively, the image of each elementary object area can be exploited to extract information about the phase and amplitude of the electron wave emerging from the specimen.

A STEM imaging mode that is capable of furnishing very high resolution images has largely superseded the modes described above. Electrons scattered through relatively wide angles (Rutherford scattering) and collected by an annular detector with appropriate dimensions form an "incoherent" image of the specimen structure, but with phase information converted into amplitude variations in the image. Atomic columns can be made visible by this technique, which is rapidly gaining importance.

The effect of partial coherence in the STEM can be analyzed by a reasoning similar to that followed for the TEM; we will not reproduce this here.

C. Image Processing

1. Interference and Holography

The resolution of electron lenses is, as we have seen, limited by the spherical aberration of the objective lens, and many types of correctors have been devised in the hope of overcoming this limit. It was realized by Dennis Gabor in the late 1940s, however, that although image detail beyond the limit cannot be discerned by eye, the information is still there if only we could retrieve it. The method he proposed for doing this was holography, but it was many years before his idea could be successfully put into practice; this had to await the invention of the laser and the development of high-performance electron microscopes. With the detailed understanding of electron image formation, the intimate connection between electron holography, electron interference, and transfer theory has become much clearer, largely thanks to Karl-Joseph Hanszen and colleagues in Braunschweig. The electron analogs of the principal holographic modes have been thoroughly explored with the aid of the Möllenstedt biprism. In the hands of Akira Tonomura in Tokyo and Hannes Lichte in Tübingen, electron holography has become a tool of practical importance.

The simplest type of hologram is the Fraunhofer in-line hologram, which is none other than a defocused electron image. Successful reconstruction requires a very coherent source (a field-emission gun) and, if the reconstruction is performed light-optically rather than digitally, glass lenses with immense spherical aberration. Such holograms should permit high-contrast detection of small weak objects.

The next degree of complexity is the single-sideband hologram, which is a defocused micrograph obtained with half of the diffraction pattern plane obscured. From the two complementary holograms obtained by obscuring each half in turn, separate phase and amplitude reconstruction is, in principle, possible. Unfortunately, this procedure is extremely difficult to put into practice, because charge accumulates along the edge of the plane that cuts off half the aperture and severely distorts the wave fronts in its vicinity; compensation is possible, but the usefulness of the technique is much reduced.

In view of these comments, it is not surprising that off-axis holography, in which a true reference wave interferes with the image wave in the recording plane, has completely supplanted these earlier arrangements. In the in-line methods, the reference wave is, of course, to be identified with the unscattered part of the main beam. Figure 28 shows an arrangement suitable for obtaining the hologram; the reference wave and image wave are recombined by the electrostatic counterpart of a biprism. In the reconstruction step, a reference wave must again be suitably combined with the wave field generated by the hologram, and the most suitable arrangement has been found to be that of the Mach–Zehnder interferometer. Many spectacular results have been obtained in this way, largely thanks to the various interference techniques developed by the school of A. Tonomura and the Bolognese group. Here, the reconstructed image is made to interfere with a plane wave. The two may be exactly aligned and yield an interference pattern representing the magnetic field in the specimen, for example; often, however, it is preferable to arrange that they are slightly inclined with respect to one another since phase "valleys" can then be distinguished from "hills." In another arrangement, the twin images are made to interfere, thereby amplifying the corresponding phase shifts twofold (or more, if higher order diffracted beams are employed).

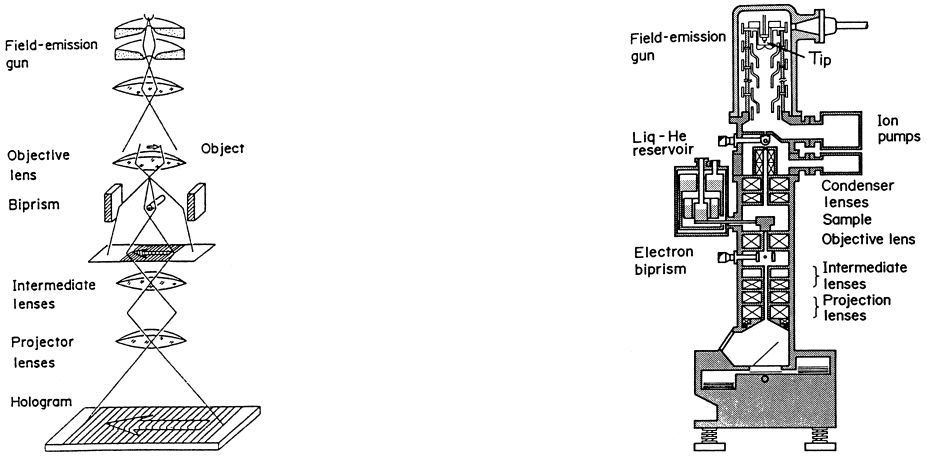


FIGURE 28 (Left) Ray diagram showing how an electron hologram is formed. (Right) Cross-section of an electron microscope equipped for holography. [From Tonomura, A. (1999). "Electron Holography," Springer-Verlag, Berlin/New York.]

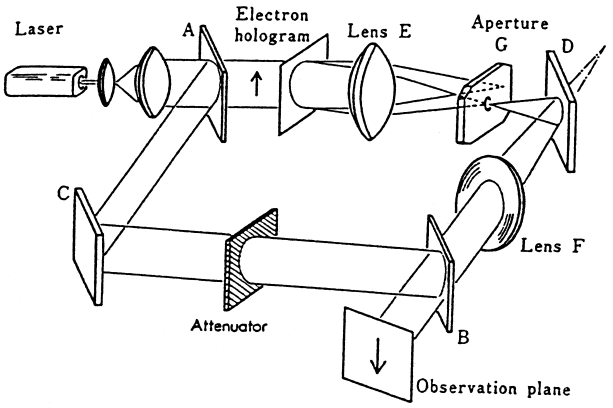


FIGURE 29 Arrangement of lenses and mirrors suitable for interference microscopy. [From Tonomura A. (1999). "Electron Holography," Springer-Verlag, Berlin/New York.]

Electron holography has a great many ramifications, which we cannot describe here, but we repeat that many of the problems that arise in the reconstruction step vanish if the hologram is available in digital form and can hence be processed in a computer. We now examine the related techniques, although not specifically in connection with holography.

2. Digital Processing

If we can sample and measure the gray levels of the electron image accurately and reliably, we can employ the computer to process the resulting matrix of image gray-level measurements in many ways. The simplest techniques, usually known as image enhancement, help to adapt the image to the visual response or to highlight features of particular interest. Many of these are routinely available on commercial scanning microscopes, and we will say no more about them here. The class of methods that allow image restoration to be achieved offer solutions of more difficult problems. Restoration filters, for example, reduce the adverse effect of the transfer functions of Eq. (81). Here, we record two or more images with different values of the defocus and hence with different forms of the transfer function and seek the weighted linear combinations of these images, or rather of their spatial frequency spectra, that yield the best estimates (in the least-squares sense) of $\tilde{\varphi}$ and $\tilde{s}$. By using a focal series of such images, we can both cancel, or at least substantially reduce, the effect of the transfer functions $\sin\gamma$ and $\cos\gamma$ and fill in the information missing from each individual picture around the zeros of these functions.

Another problem of considerable interest, in other fields as well as in electron optics, concerns the phase of the object wave for strongly scattering objects. We have seen that the specimens studied in transmission microscopy are essentially transparent: The image is formed not by absorption but by scattering. The information about the specimen is therefore in some sense coded in the angular distribution of the electron trajectories emerging from the specimen. In an ideal system, this angular distribution would be preserved, apart from magnification effects, at the image and no contrast would be seen. Fortunately, however, the microscope is imperfect; contrast is generated by the loss of electrons scattered through large angles and intercepted by the diaphragm or "objective aperture" and by the crude analog of a phase plate provided by the combination of spherical aberration and defocus. It is the fact that the latter affects the angular distribution within the beam and converts it to a positional dependence with a fidelity that is measured by the transfer function $\sin\gamma$ that is important. The resulting contrast can be related simply to the specimen transparency only if the phase and amplitude variations are small, however, and this is true of only a tiny class of specimens. For many of the remainder, the problem remains. It can be expressed graphically by saying that we know from our intensity record where the electrons arrive (amplitude) but not their directions of motion at the point of arrival (phase). Several ways of obtaining this missing information have been proposed, many inspired by the earliest suggestion, the Gerchberg–Saxton algorithm. Here, the image and diffraction pattern of exactly the same area are recorded, and the fact that the corresponding wave functions are related by a Fourier transform is used to find the phase iteratively. First, the known amplitudes in the image, say, are given arbitrary phases and the Fourier transform is calculated; the amplitudes thus found are then replaced by the known diffraction pattern amplitudes and the process is repeated. After several iterations, the unknown phases should be recovered. This procedure encounters many practical difficulties and some theoretical ones as well, since the effect of noise is difficult to incorporate. This and several related algorithms have now been thoroughly studied and their reliability is well understood. In these iterative procedures, two signals generated by the object are required (image and diffraction pattern or two images at different defocus values in particular). If a STEM is used, this multiplicity of information is available in a single record if the intensity distribution associated with every object pixel is recorded and not reduced to one or a few summed values. A sequence of Fourier transforms and mask operations that generate the phase and amplitude of the electron wave has been devised by John Rodenburg.

A very different group of methods has grown up around the problem of three-dimensional reconstruction. The two-dimensional projected image that we see in the microscope often gives very little idea of the complex spatial relationships of the true structure, and techniques have therefore

been developed for reconstructing the latter. They consist essentially in combining the information provided by several different views of the specimen, supplemented if possible by prior knowledge of an intrinsic symmetry of the structure. The fact that several views are required reminds us that not all specimens can withstand the electron onslaught that such multiple exposure represents. Indeed, there are interesting specimens that cannot be directly observed at all, because they are destroyed by the electron dose that would be needed to form a discernible image. Very low dose imaging must therefore be employed, and this has led to the development of an additional class of image restoration methods. Here, the aim is first to detect the structures, invisible to the unaided eye, and superimpose low-dose images of identical structures in such a way that the signal increases more rapidly than the noise and so gradually emerges from the surrounding fog. Three-dimensional reconstruction may then be the next step. The problem here, therefore, is first to find the structures, then to align them in position and orientation with the precision needed to achieve the desired resolution. Some statistical check must be applied to be sure that all the structures found are indeed the same and not members of distinct groups that bear a resemblance to one other but are not identical. Finally, individual members of the same group are superposed. Each step demands a different treatment. The individual structures are first found by elaborate cross-correlation calculations. Cross-correlation likewise enables us to align them with high precision. Multivariate analysis is then used to classify them into groups or to prove that they do, after all, belong to the same group and, a very important point, to assign probabilities to their membership of a particular group.

IV. CONCLUDING REMARKS

Charged-particle optics has never remained stationary with the times, but the greatest upheaval has certainly been that caused by the widespread availability of large, fast computers. Before, the analysis of electron lenses relied heavily on rather simple field or potential models, and much ingenuity was devoted to finding models that were at once physically realistic and mathematically tractable. Apart from sets of measurements, guns were almost virgin territory. The analysis of in-lens deflectors would have been unthinkable but fortunately was not indispensable since even the word *microminiaturization* has not yet been coined. Today, it is possible to predict with great accuracy the behavior of almost any system; it is even possible to obtain aberration coefficients, not by evaluating the corresponding integrals, themselves obtained as a result of exceedingly long and tedious algebra, but by solving the exact ray equations and fitting the results to the known aberration pattern. This is particularly valuable when parasitic aberrations, for which aberration integrals are not much help, are being studied. Moreover, the aberration integrals can themselves now be established not by long hours of laborious calculation, but by means of one of the computer algebra languages. A knowledge of the fundamentals of the subject, presented here, will always be necessary for students of the subject, but modern numerical methods now allow them to go as deeply as they wish into the properties of the most complex systems.

SEE ALSO THE FOLLOWING ARTICLES

ACCELERATOR PHYSICS AND ENGINEERING • HOLOGRAPHY • QUANTUM OPTICS • SCANNING ELECTRON MICROSCOPY • SCATTERING AND RECOILING SPECTROSCOPY • SIGNAL PROCESSING, DIGITAL • WAVE PHENOMENA

BIBLIOGRAPHY

Carey, D. C. (1987). "The Optics of Charged Particle, Beams," Harwood Academic, London.

Chapman, J. N., and Craven, A. J., eds. (1984). "Quantitative Electron Microscopy," SUSSP, Edinburgh.

Dragt, A. J., and Forest, E. (1986). *Adv. Electron. Electron. Phys.* **67,** 65–120.

Feinerman, A. D. and Crewe, D. A. (1998). "Miniature electron optics." *Adv. Imaging Electron Phys*. **102,** 187–234.

Frank, J. (1996). "Three-Dimensional Electron Microscopy of Macromolecular Assemblies," Academic Press, San Diego.

Glaser, W. (1952). "Grundlagen der Elektronenoptik," Springer-Verlag, Vienna.

Glaser, W. (1956). Elektronen- und Ionenoptik, *Handb. Phys.* **33,** 123–395.

Grivet, P. (1972). "Electron Optics," 2nd Ed. Pergamon, Oxford.

Hawkes, P. W. (1970). Adv. Electron. Electron Phys., Suppl. 7. Academic Press, New York.

Hawkes, P. W., ed. (1973). "Image Processing and Computer-Aided Design in Electron Optics," Academic Press, New York.

Hawkes, P. W., ed. (1980). "Computer Processing of Electron Microscope Images," Springer-Verlag, Berlin and New York.

Hawkes, P. W., ed. (1982). "Magnetic Electron Lenses," Springer-Verlag, Berlin and New York.

Hawkes, P. W., and Kasper, E. (1989, 1994). "Principles of Electron Optics," Academic Press, San Diego.

Hawkes, P. W., ed. (1994). "Selected Papers on Electron Optics," SPIE Milestones Series, Vol. 94.

Humphries, S. (1986). "Principles of Charged Particle Acceleration." (1990). "Charged Particle Beams," Wiley-Interscience, New York and Chichester.

Lawson, J. D. (1988). "The Physics of Charged-Particle Beams," Oxford Univ. Press, Oxford.

Lencová, B. (1997). Electrostatic Lenses. In "Handbook of Charged Particle Optics" (J. Orloff, ed.), pp. 177–221, CRC Press, Boca Raton, FL.

Livingood, J. J. (1969). "The Optics of Dipole Magnets," Academic Press, New York.
Orloff, J., ed. (1997). "Handbook of Charged Particle Optics," CRC Press, Boca Raton, FL.
Reimer, L. (1997). "Transmission Electron Microscopy," Springer-Verlag, Berlin and New York.
Reimer, L. (1998). "Scanning Electron Microscopy," Springer-Verlag, Berlin and New York.
Saxton, W. O. (1978). Adv. Electron. Electron Phys., Suppl. 10. Academic Press, New York.
Septier, A., ed. (1967). "Focusing of Charged Particles," Vols. 1 and 2, Academic Press, New York.
Septier, A., ed. (1980–1983). *Adv. Electron. Electron Phys., Suppl.* **13A–C**. Academic Press, New York.
Tonomura, A. (1999). "Electron Holography," Springer-Verlag, Berlin.
Tsuno, K. (1997). Magnetic Lenses for Electron Microscopy. In "Handbook of Charged Particle Optics" (J. Orloff, ed.), pp. 143–175, CRC Press, Boca Raton, FL.
Wollnik, H. (1987). "Optics of Charged Particles," Academic Press, Orlando.



Elasticity

Herbert Reismann
State University of New York at Buffalo

I. One-Dimensional Considerations
II. Stress
III. Strain
IV. Hooke's Law and Its Limits
V. Strain Energy
VI. Equilibrium and the Formulation of Boundary Value Problems
VII. Examples

GLOSSARY

Anisotropy A medium is said to be anisotropic if the value of a measured, physical field quantity depends on the orientation (or direction) of measurement.

Eigenvalue and eigenvector Consider the matrix equation $\mathbf{AX} = \lambda\mathbf{X}$, where $\mathbf{A}$ is an $n \times n$ square matrix, and $\mathbf{X}$ is an n-dimensional column vector. In this case, the scalar λ is an eigenvalue, and $\mathbf{X}$ is the associated eigenvector.

Isotropy A medium is said to be isotropic if the value of a measured, physical field quantity is independent of orientation.

ELASTICITY THEORY is the (mathematical) study of the behavior of those solids that have the property of recovering their size and shape when the forces that cause the deformation are removed. To some extent, almost all solids display this property. In this article, most of the discussion will be limited to the special case of linearly elastic solids, where deformation is proportional to applied forces. This topic is usually referred to as classical elasticity theory. This branch of mathematical physics was formulated during the nineteenth century and, since its inception, has been developed and refined to form the background and foundation for disciplines such as structural mechanics; stress analysis; strength of materials; plates and shells; solid mechanics; and wave propagation and vibrations in solids. These topics are fundamental to solving present-day problems in many branches of modern engineering and applied science. They are used by structural (civil) engineers, aerospace engineers, mechanical engineers, geophysicists, geologists, and bioengineers, to name a few. The deformation, vibrations, and structural integrity of modern high-rise buildings, airplanes, and high-speed rotating machinery are predicted by applying the modern theory of elasticity.

I. ONE-DIMENSIONAL CONSIDERATIONS

If we consider a suitably prepared rod of mild steel, with (original) length L and cross-sectional area A, subjected

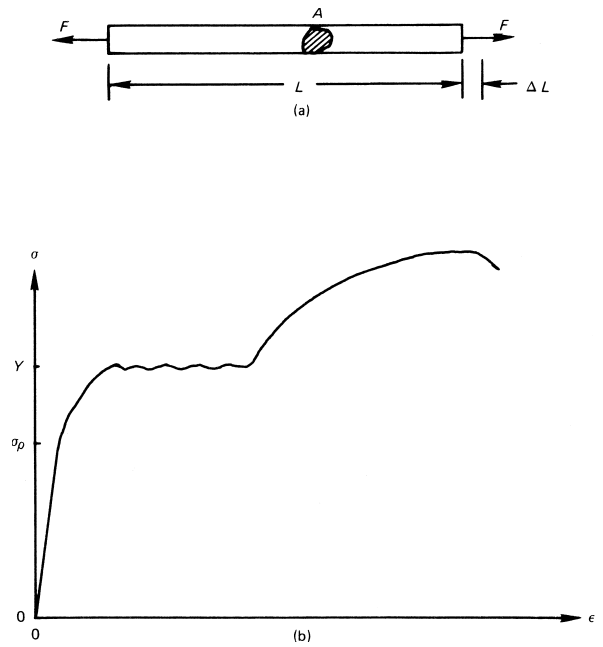


FIGURE 1 (a) Tension rod. (b) Stress–strain curve (ductile material).

to a longitudinal, tensile force of magnitude F, then the rod will experience an elongation of magnitude ΔL, as shown in Fig. la. So that we can compare the behavior of rods of differing cross section in a meaningful manner, it is convenient to define the (uniform) axial stress in the rod by $\sigma = F/A$ and the (uniform) axial strain by $\varepsilon = \Delta L/L$. We note that the unit of stress is force per unit of (original) area and the unit of strain is change in length divided by original length. If, in a typical tensile test, we plot stress σ versus strain ε, we obtain the curve shown in Fig. 1b. In the case of mild steel, and many other ductile materials, this curve has a straight line portion that extends from $0 < \sigma < \sigma_p$, where σ_p is the proportional limit. The slope of this line is $\sigma/\varepsilon = E$, where E is known as Young's modulus (Thomas Young, 1773–1829). When $\sigma_p < \sigma$, the stress–strain curve is no longer linear, as shown in Fig. 1b. When the rod is extended beyond $\sigma = \sigma_p$ (the proportional limit), it suffers a permanent set (deformation) upon removal of the load F. At $\sigma = Y$ (the yield point), the strain will increase considerably for relatively small increases in stress (Fig. 1b). For the majority of structural applications, it is desirable to remain in the linearly elastic, shape-recoverable range of stress and strain ($0 \leq \sigma \leq \sigma_p$). The mathematical material model that is based on this assumption is said to display linear material characteristics. For example, an airplane wing will deflect in flight because of air loads and maneuvers, but when the loads are removed, the wing reverts to its original shape. If this were not the case, the wing's lifting capability would not be reliably predictable, and, of course, this would not be desirable. In addition, if the load is doubled, the deflection will also double.

Within the context of the international system of units (*Système International*, or SI), the unit of stress is the pascal (Pa). One pascal is equal to one newton per square meter (N m^{-2}). The unit of strain is meter per meter, and thus strain is a dimensionless quantity. We note that 1 N m^{-2} = 1 Pa = 1.4504×10^{-4} psi and 1 psi = 6894.76 Pa.

Typical values of the Young's (elastic) modulus E and yield stress in tension Y for some ductile materials are shown in Table II in Section IV. The tension test of a rod and naive definitions of stress and strain are associated with one-dimensional considerations. Elasticity theory is concerned with the generalization of these concepts to the general, three-dimensional case.

II. STRESS

Elastic solids are capable of transmitting forces, and the concept of stress in a solid is a sophistication and generalization of the concept of force. We consider a material point P in the interior of an elastic solid and pass an oriented plane II through P with unit normal vector $\mathbf{n}$ (see Fig. 2). Consider the portion of the solid which is shaded. Then on a (small) area ΔA surrounding the point P, there will act a net force of magnitude $\Delta\mathbf{F}$, and the stress vector at P is defined by the limiting process

$$\mathbf{T}(\mathbf{n}) = \lim_{\Delta A \to 0} \frac{\Delta \mathbf{F}}{\Delta A}. \tag{1}$$

It is to be noted that the magnitude as well as the direction of the stress vector $\mathbf{T}$ depends upon the orientation of $\mathbf{n}$. If we resolve the stress vector along the (arbitrarily chosen) $(x, y, z) = (x_1, x_2, x_3)$ axes, then we can write

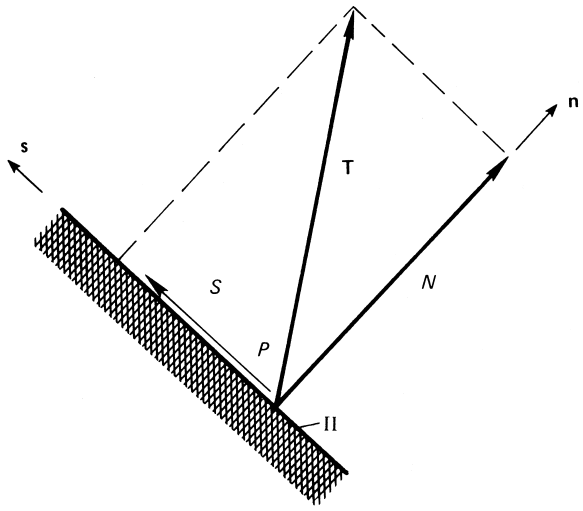


FIGURE 2 Stress vector and components.

$$\begin{aligned}\mathbf{T}_1 &= \tau_{11}\mathbf{e}_1 + \tau_{12}\mathbf{e}_2 + \tau_{13}\mathbf{e}_3\\ \mathbf{T}_2 &= \tau_{21}\mathbf{e}_1 + \tau_{22}\mathbf{e}_2 + \tau_{23}\mathbf{e}_3\\ \mathbf{T}_3 &= \tau_{31}\mathbf{e}_1 + \tau_{32}\mathbf{e}_2 + \tau_{33}\mathbf{e}_3,\end{aligned} \tag{2}$$

where $\mathbf{T}_i = \mathbf{T}(\mathbf{e}_i)$ for $i = 1, 2, 3$; that is, the $\mathbf{T}_i$ are stress vectors acting upon the three coordinate planes and $\mathbf{e}_i$ are unit vectors associated with the coordinate axes $(x, y, z) = (x_1, x_2, x_3)$. We note that here and in subsequent developments, we use the convenient and common notation $\tau_{12} \equiv \tau_{xy}, \mathrm{T}_1 \equiv \mathrm{T}_x, \mathrm{T}_2 \equiv \mathrm{T}_y$, etc. In other words, the subscripts 1, 2, 3 take the place of x, y, z. We can also write

$$\begin{aligned}T_1 &= \tau_{11}n_1 + \tau_{12}n_2 + \tau_{13}n_3\\ T_2 &= \tau_{21}n_1 + \tau_{22}n_2 + \tau_{23}n_3\\ T_3 &= \tau_{31}n_1 + \tau_{32}n_2 + \tau_{33}n_3,\end{aligned} \tag{3}$$

where

$$\mathbf{n} = \mathbf{e}_1 n_1 + \mathbf{e}_2 n_2 + \mathbf{e}_3 n_3$$

and

$$\begin{aligned}\mathbf{T}(\mathbf{n}) &= \mathbf{e}_1 T_1 + \mathbf{e}_2 T_2 + \mathbf{e}_3 T_3\\ &= \mathbf{T}_1 n_1 + \mathbf{T}_2 n_2 + \mathbf{T}_3 n_3.\end{aligned} \tag{4}$$

This last expression is known as the lemma of Cauchy (A. L. Cauchy, 1789–1857). The stress tensor components

$$[\tau_{ij}] = \begin{bmatrix}\tau_{11} & \tau_{12} & \tau_{13}\\ \tau_{21} & \tau_{22} & \tau_{23}\\ \tau_{31} & \tau_{32} & \tau_{33}\end{bmatrix} \tag{5}$$

can be visualized with reference to Fig. 3, with all stresses shown acting in the positive sense. We note that τ_{ij} is the stress component acting on the face with normal $\mathbf{e}_i$, in the direction of the vector $\mathbf{e}_j$.

With reference to Fig. 2, it can also be shown that relative to the plane II, the normal component N and the shear component S are given by

$$\begin{aligned}T_n \equiv N &= \mathbf{T}\cdot\mathbf{n} = \sum_{i=1}^{3} T_i n_i\\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} n_i n_j \tau_{ij}\end{aligned} \tag{6a}$$

and

$$\begin{aligned}T_s \equiv S &= \mathbf{T}\cdot\mathbf{s} = \sum_{i=1}^{3} T_i s_i\\ &= \sum_{i=1}^{3}\sum_{j=1}^{3} n_i s_j \tau_{ij},\end{aligned} \tag{6b}$$

where $\mathbf{n}\cdot\mathbf{s} = 0$ and $\mathbf{n}, \mathbf{s}$ are unit vectors normal and parallel to the plane II, respectively.

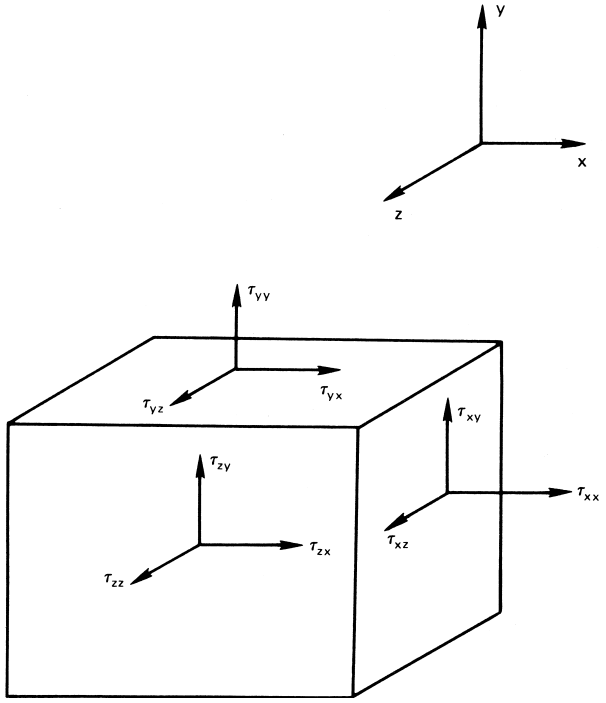


FIGURE 3 Stress tensor components.

At every interior point of a stressed solid, there exist at least three mutually perpendicular directions for which all shearing stresses $\tau_{ij}, i \neq j$, vanish. This preferred axis system is called the principal axis system. It can be found by solving the algebraic eigenvalue–eigenvector problem characterized by

$$\begin{bmatrix}\tau_{11}-\sigma & \tau_{12} & \tau_{13}\\ \tau_{21} & \tau_{22}-\sigma & \tau_{23}\\ \tau_{31} & \tau_{32} & \tau_{33}-\sigma\end{bmatrix}\begin{Bmatrix}n_1\\ n_2\\ n_3\end{Bmatrix} = \begin{Bmatrix}0\\ 0\\ 0\end{Bmatrix}, \tag{7}$$

where n, n_2, and n_3 are the direction cosines of the principal axis system such that $n_1^2 + n_2^2 + n_3^2 = 1$; and σ_1, σ_2, and σ_3 are the (scalar) principal stress components. The necessary and sufficient condition for the existence of a solution for Eq. (7) is obtained by setting the coefficient determinant equal to zero. The result is

$$\sigma^3 + I_1\sigma^2 + I_2\sigma - I_3 = 0, \tag{8}$$

where the quantities

$$I_1 = \tau_{11} + \tau_{22} + \tau_{33}, \tag{9a}$$

$$I_2 = \begin{vmatrix}\tau_{11} & \tau_{12}\\ \tau_{21} & \tau_{22}\end{vmatrix} + \begin{vmatrix}\tau_{22} & \tau_{23}\\ \tau_{32} & \tau_{33}\end{vmatrix} + \begin{vmatrix}\tau_{33} & \tau_{31}\\ \tau_{13} & \tau_{11}\end{vmatrix}, \tag{9b}$$

and

$$I_3 = \begin{vmatrix}\tau_{11} & \tau_{12} & \tau_{13}\\ \tau_{21} & \tau_{22} & \tau_{23}\\ \tau_{31} & \tau_{32} & \tau_{33}\end{vmatrix} \tag{9c}$$

are known as the first, second, and third stress invariants, respectively. For example, we consider these stress tensor components at a point P of a solid, relative to the x, y, z axes:

$$[\tau_{ij}] = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{bmatrix}. \tag{10}$$

Thus,

$$I_1 = 3, \qquad I_2 = -6, \qquad I_3 = -8$$

and

$$\sigma^3 - 3\sigma^2 - 6\sigma + 8 = (\sigma - 4)(\sigma - 1)(\sigma + 2) = 0.$$

Consequently, the principal stresses at P are $\sigma_1 = 4$, $\sigma_2 = 1$, and $\sigma_3 = -2$. With the aid of Eq. (7), it can be shown that the principal directions at P are given by the mutually perpendicular unit vectors

$$\begin{aligned} \mathbf{n}^{(1)} &= \mathbf{e}_1 \frac{2}{\sqrt{6}} + \mathbf{e}_2 \frac{1}{\sqrt{6}} + \mathbf{e}_3 \frac{1}{\sqrt{6}} \\ \mathbf{n}^{(2)} &= \mathbf{e}_1 \left(-\frac{1}{\sqrt{3}} \right) + \mathbf{e}_2 \frac{1}{\sqrt{3}} + \mathbf{e}_3 \frac{1}{\sqrt{3}} \\ \mathbf{n}^{(3)} &= \mathbf{e}_1 (0) + \mathbf{e}_2 \left(-\frac{1}{\sqrt{2}} \right) + \mathbf{e}_3 \left(\frac{1}{\sqrt{2}} \right). \end{aligned} \tag{11}$$

When the Cartesian axes are rotated in a rigid manner from x_1, x_2, x_3, to $x_{1'}, x_{2'}, x_{3'}$, as shown in Fig. 4, the components of the stress tensor transform according to the rule

$$\tau_{p'q'} = \sum_{i=1}^{3} \sum_{j=1}^{3} a_{p'i} a_{q'j} \tau_{ij}, \tag{12}$$

where $a_{p'i} = \cos(x_{p'}, x_1) = \cos(\mathbf{e}_{p'}, \mathbf{e}_i)$ are the nine direction cosines that orient the primed coordinate system relative to the unprimed system.

For example, consider the rotation of axes characterized by the table of direction cosines

$$[a_{i'j}] = \begin{bmatrix} \frac{2}{3} & -\frac{2}{3} & -\frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \end{bmatrix}. \tag{13}$$

The stress components τ_{ij} in Eq. (10) relative to the x, y, z axes will become

$$[\tau_{p'q'}] = \begin{bmatrix} 0.889 & 0.778 & 0.222 \\ 0.778 & -1.444 & 1.444 \\ 0.222 & 1.444 & 3.556 \end{bmatrix} \tag{14}$$

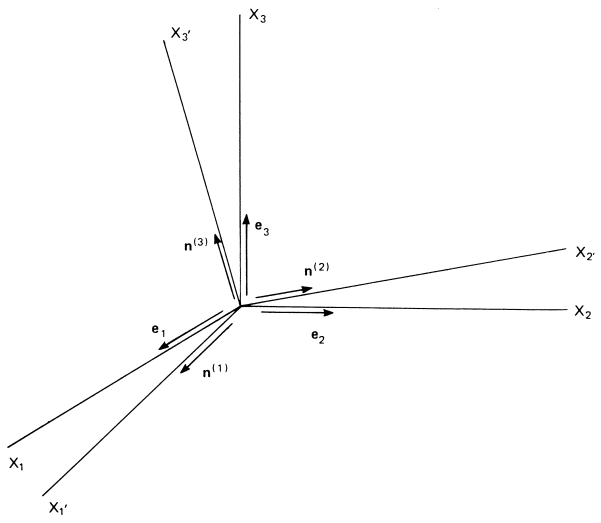


FIGURE 4 Principal axes.

when referred to x', y', z' axes, according to the law of transformation [Eq. (12)]. The extreme shear stress at a point is given by $\tau_{\max} = \frac{1}{2}(\sigma_1 - \sigma_3)$ and this value is $\tau_{\max} = \frac{1}{2}(4 + 2) = 3$ for the stress tensor [Eq. (10)]. It should be noted that the principal stresses are ordered, that is, $\sigma_1 \geq \sigma_2 \geq \sigma_3$, and that $\sigma_1(\sigma_3)$ is the largest (smallest) normal stress for all possible planes through the point P.

If we now establish a coordinate system coincident with principal axes then in "principal stress space," the normal stress N and the shear stress S on a plane characterized by the outer unit normal vector $\mathbf{n}$ are, respectively,

$$N = n_1^2 \sigma_1 + n_2^2 \sigma_2 + n_3^2 \sigma_3 \tag{15a}$$

and

$$S^2 = n_1^2 n_2^2 (\sigma_1 - \sigma_2)^2 + n_2^2 n_3^2 (\sigma_2 - \sigma_3)^2 + n_3^2 n_1^2 (\sigma - \sigma_1)^2, \tag{15b}$$

where σ_1, σ_2, and σ_3 are principal stresses. We now visualize eight planes, the normal to each of which makes equal angles with respect to principal axes. The shear stress acting upon these planes is known as the octahedral shear stress τ_0, and its magnitude is

$$\tau_0 = \tfrac{1}{3}\left[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 \right]^{1/2} \geq 0. \tag{16}$$

It can be shown that the octahedral shear stress is related to the average of the square of all possible shear stresses at the point, and the relation is

$$\tfrac{3}{5}(\tau_0)^2 = \langle S^2 \rangle. \tag{17}$$

It can also be shown that

$$9\tau_0^2 = 2I_1^2 - 6I_2, \tag{18}$$

where I_1 and I_2 are the first and second stress invariants, respectively [see Eqs. (9a) and (9b)]. We also note the bound

$$1 \le \sqrt{\frac{3}{2}}\frac{\tau_0}{\tau_{\max}} \le \frac{2}{\sqrt{3}} \tag{19}$$

and the associated implication that $\frac{3}{2}\tau_0 \cong 1.08\ \tau_{\max}$ with a maximum error of about 7%. Returning to the stress tensor [Eq. (10)], we have

$$\tau_{\max} = 3$$

and

$$9\tau_0^2 = 2I_1^2 - 6I_2 = (2)(9) + (6)(6) = 54,$$

or

$$\tau_0 = \sqrt{6} = 2.4495,$$

and

$$1 \le \frac{\sqrt{3/2}\tau_0}{\tau_{\max}} = \frac{\sqrt{3/2}(2.4495)}{3} \le \frac{2}{\sqrt{3}},$$

or

$$1 = 1 \le 1.1547.$$

III. STRAIN

In our discussion of the concept of stress, we noted that stress characterizes the action of a "force at a point" in a solid. In a similar manner, we shall show that the concept of strain can be used to quantify the notion of "deformation at a point" in a solid.

We consider a (small) quadrilateral element in the unstrained solid with dimensions dx, dy, and dz. The sides of the element are taken to be parallel to the coordinate axes. After deformation, the volume element has the shape of a rectangular parallelepiped with edges of length $(dx+du)$, $(dy+dv)$, $(dz+dw)$. With reference to Fig. 5, the material point P in the undeformed configuration is carried into the point P' in the deformed configuration. A projection of the element sides onto the x–y plane, before and after deformation, is shown in Fig. 5. We note that all changes in length and angles are small, and they

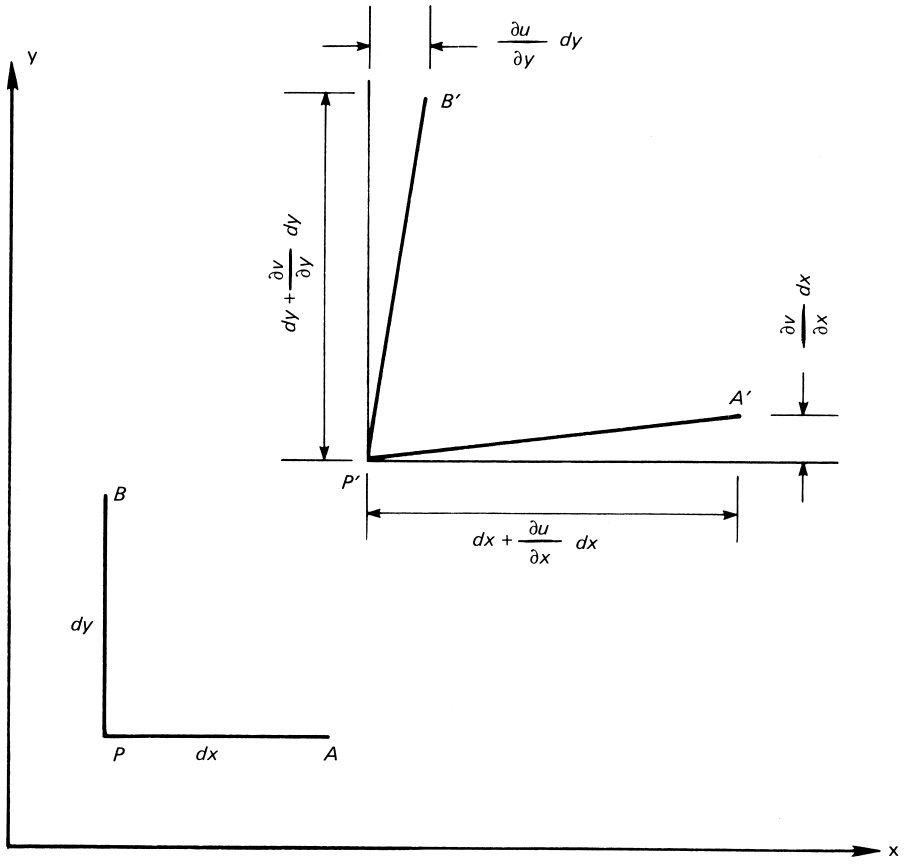


FIGURE 5 Strain.

have been exaggerated for purposes of clarity. We now define extensional strain $\varepsilon_{xx} = \varepsilon_{11}$ as change in length per unit length, and therefore for the edge PA (in Fig. 5), we have

$$\varepsilon_{xx} = \frac{[dx + (\partial u/\partial x)\,dx] - dx}{dx} = \frac{\partial u}{\partial x}$$

and

$$\varepsilon_{yy} = \frac{[dx + (\partial v/\partial y)\,dy] - dy}{dy} = \frac{\partial v}{\partial y},$$

and a projection onto the y–z plane will result in

$$\varepsilon_{zz} = \frac{[dz + (\partial w/\partial z)\,dz] - dz}{dz} = \frac{\partial w}{\partial z}.$$

The shear strain is defined as one-half of the decrease of the originally right angle APB. Thus, with reference to Fig. 5, we have

$$2\varepsilon_{xy} = 2\varepsilon_{yx} = \frac{(\partial v/\partial x)\,dx}{dx + (\partial u/\partial x)\,dx} + \frac{(\partial u/\partial y)\,dy}{dy + (\partial v/\partial y)\,dy}$$

$$= \frac{\partial v/\partial x}{1 + (\partial u/\partial x)} + \frac{\partial u/\partial y}{1 + (\partial v/\partial y)} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}$$

because it is assumed that

$$1 \gg \partial u/\partial x; \qquad 1 \gg \partial v/\partial y \qquad \text{(small rotations)}.$$

In a similar manner, using projections onto the planes y–z and z–x, we can show that

$$2\varepsilon_{yz} = \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y}$$

and

$$2\varepsilon_{zx} = \frac{\partial w}{\partial x} + \frac{\partial u}{\partial z}.$$

Consequently, the complete (linearized) strain-displacement relations are given by

$$\begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{yx} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{zx} & \varepsilon_{zy} & \varepsilon_{zz} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\partial u}{\partial x} & \dfrac{1}{2}\left(\dfrac{\partial u}{\partial y} + \dfrac{\partial v}{\partial x}\right) & \dfrac{1}{2}\left(\dfrac{\partial u}{\partial z} + \dfrac{\partial w}{\partial x}\right) \\ \dfrac{1}{2}\left(\dfrac{\partial v}{\partial x} + \dfrac{\partial u}{\partial y}\right) & \dfrac{\partial v}{\partial y} & \dfrac{1}{2}\left(\dfrac{\partial v}{\partial z} + \dfrac{\partial w}{\partial y}\right) \\ \dfrac{1}{2}\left(\dfrac{\partial w}{\partial x} + \dfrac{\partial u}{\partial z}\right) & \dfrac{1}{2}\left(\dfrac{\partial w}{\partial y} + \dfrac{\partial v}{\partial z}\right) & \dfrac{\partial w}{\partial z} \end{bmatrix}. \tag{20}$$

Equation (20) characterizes the deformation of the solid at a point. If we define the mutually perpendicular unit vectors **n** and **s** with reference to a plane II through a point P in a solid (see Fig. 2), then it can be shown that the extensional strain N in the direction **n** is given by the formula

$$N = \sum_{i=1}^{3}\sum_{j=1}^{3} \varepsilon_{ij} n_i n_j \tag{21a}$$

and the shear strain relative to the vectors **n** and **s** is

$$S = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3} \varepsilon_{ij} n_i s_j. \tag{21b}$$

Equation (21a) expresses the extensional strain and Eq. (21b) expresses the shearing strain for an arbitrarily chosen element; therefore, we can infer that the nine (six independent) quantities $\varepsilon_{ij} (i = 1, 2, 3; j = 1, 2, 3)$ provide a complete characterization of strain associated with a material point in the solid. It can be shown that the nine quantities ε_{ij} constitute the components of a tensor of order two in a three-dimensional space, and the appropriate law of transformation under a rotation of coordinate axes is

$$\varepsilon_{p'q'} = \sum_{i=1}^{3}\sum_{j=1}^{3} a_{p'i} a_{q'j} \varepsilon_{ij}; \tag{22}$$

$$p = 1, 2, 3; \qquad q = 1, 2, 3,$$

where the $a_{p'i}$ are direction cosines as in Eq. (12). As in the case of stress, there will be at least one set of mutually perpendicular axes for which the shearing strains vanish. These axes are principal axes of strain. They are found in a manner that is entirely analogous to the determination of principal stresses and axes. (See Section II.)

It should be noted that a single-valued, continuous displacement field for a simply connected region is guaranteed provided that the six equations of compatibility of A. J. C. Barré de Saint-Venant (1779–1886) are satisfied:

$$\frac{\partial^2 \varepsilon_{xx}}{\partial y^2} + \frac{\partial^2 \varepsilon_{yy}}{\partial x^2} = 2\frac{\partial^2 \varepsilon_{xy}}{\partial x \partial y}, \tag{23a}$$

$$\frac{\partial^2 \varepsilon_{xx}}{\partial y \partial z} = \frac{\partial}{\partial x}\left(-\frac{\partial e_{yx}}{\partial x} + \frac{\partial \varepsilon_{xz}}{\partial y} + \frac{\partial \varepsilon_{xy}}{\partial z}\right), \tag{23b}$$

and there are two additional equations for each of Eqs. (23a) and (23b), which are readily obtained by cyclic permutation of x, y, z.

IV. HOOKE'S LAW AND ITS LIMITS

The most general linear relationship between stress tensor and strain tensor components at a point in a solid is given by

$$\tau_{ij} = \sum_{k=1}^{3}\sum_{l=1}^{3} C_{ijkl}\varepsilon_{kl}; \qquad i = 1, 2, 3; \quad j = 1, 2, 3, \tag{24}$$

where the $3^4 = 81$ constants C_{ijkl} are the elastic constants of the solid. If a strain energy density function exists (see Section V), and in view of the fact that the stress and strain tensor components are symmetric, the elastic constants must satisfy the relations

$$C_{ijkl} = C_{ijlk}, \qquad C_{ijkl} = C_{jikl}, \qquad C_{ijkl} = C_{klij}, \tag{25}$$

and therefore the number of independent elastic constants is reduced to $\frac{1}{2}(6^2 - 6) + 6 = 21$ for the general anisotropic elastic solid. If, in addition, the elastic properties of the solid are independent of orientation, the number of independent elastic constants can be reduced to two. In this case of an isotropic elastic solid, the relation between stress and strain is given by

$$\begin{aligned} E\varepsilon_{xx} &= \tau_{xx} - \nu(\tau_{yy} + \tau_{zz}) \\ E\varepsilon_{yy} &= \tau_{yy} - \nu(\tau_{zz} + \tau_{xx}) \\ E\varepsilon_{zz} &= \tau_{zz} - \nu(\tau_{xx} + \tau_{yy}) \\ 2G\varepsilon_{xy} &= \tau_{xy} \\ 2G\varepsilon_{yz} &= \tau_{yz} \\ 2G\varepsilon_{zx} &= \tau_{zx}, \end{aligned} \tag{26}$$

where $G = E/2(1 + \nu)$ is the shear modulus, E is Young's modulus (see Section I), and ν is Pohisson's ratio (S. D. Poisson, 1781–1840). Equation (26) is known as Hooke's law (Robert Hooke, 1635–1693) for a linearly elastic, isotropic solid. A listing of typical values of the elastic constants is provided in Table I.

Many failure theories for solids have been proposed, and they are usually associated with specific classes of materials. In the case of a ductile material with a well-defined yield point (see Fig. 1b), there are at least two failure theories that yield useful results.

TABLE I Typical Values of Elastic Constants[a]

Material	ν	E (Pa)[b]	G (Pa)[b]
Aluminum	0.34	6.89×10^{10}	2.57×10^{10}
Concrete	0.20	0.76×10^{10}	1.15×10^{10}
Copper	0.34	8.96×10^{10}	3.34×10^{10}
Glass	0.25	6.89×10^{10}	2.76×10^{10}
Nylon	0.40	2.83×10^{10}	1.01×10^{10}
Rubber	0.499	1.96×10^{6}	0.654×10^{6}
Steel	0.29	20.7×10^{10}	8.02×10^{10}

[a] Adapted from Reismann, H., and Pawlik, P. S. (1980). "Elasticity: Theory and Applications," Wiley (Interscience), New York.
[b] Note that 1 Pa = 1 N m^{-2} = 1.4504 $\times 10^{-4}$ lb in.$^{-2}$.

TABLE II Some Material Properties for Ductile Materials[a]

Material	Yield point stress, σ_Y (tension, Pa)	Young's modulus, E (Pa)	Strain at yield point, ε_Y (tension)
Aluminum alloy (2024 T 4)	290×10^{6}	7.30×10^{10}	0.00397
Brass	103×10^{6}	10.3×10^{10}	0.00100
Bronze	138×10^{6}	10.3×10^{10}	0.00134
Magnesium alloy	138×10^{6}	4.50×10^{10}	0.00307
Steel (low carbon, structural	248×10^{6}	20.7×10^{10}	0.00120
Steel (high carbon)	414×10^{6}	20.7×10^{10}	0.00200

[a] Adapted from Reismann, H., and Pawlik, P. S. (1980). "Elasticity: Theory and Applications," Wiley (Interscience), New York.

A. The Hencky-Mises Yield Criterion

This theory predicts failure (yielding) at a point of the solid when $9\tau_0^2 \geq 2Y^2$, where τ_0 is the octahedral shear stress [see Eq. (16)] and Y is the yield stress in tension (see Fig. 1b). In this case, the ratio of yield stress in tension Y to the yield stress in pure shear τ has the value $Y/\tau = \sqrt{3}$.

B. The Tresca Yield Criterion

This theory postulates that yielding occurs when the extreme shear stress $\tau_{\max}$ at a point attains the value $\tau_{\max} \geq Y/2$. We note that for this theory the ratio of yield stress in tension to the yield stress in pure shear is equal to $Y/\tau = 2$. A listing of the values of Y for some commonly used materials is given in Table II.

V. STRAIN ENERGY

We now consider an interior material point P in a stressed, elastic solid. We can construct a Cartesian coordinate system x, y, z with origin at P, which is coincident with principal axes at P. The point P is enclosed by a small, rectangular parallelepiped with sides of length dx, dy, and dz. The areas of the sides of the parallelepiped are $dA_z = dx\,dy$, $dA_x = dy\,dz$, $dA_y = dz\,dx$, and the volume is $dV = dx\,dy\,dz$. The potential (or strain) energy stored in the linearly elastic solid is equal to the work of the external forces. Consequently, neglecting heat generation, if W is the strain energy per unit volume (strain energy density), we have

$$
\begin{aligned}
W\,dV &= \tfrac{1}{2}(\tau_{xx}A_x)(dx\varepsilon_{xx}) + \tfrac{1}{2}(\tau_{yy}A_y)(dy\varepsilon_{yy}) \\
&\quad + \tfrac{1}{2}(\tau_{zz}A_z)(dz\varepsilon_{zz}) \\
&= \tfrac{1}{2}(\tau_{xx}\varepsilon_{xx} + \tau_{yy}\varepsilon_{yy} + \tau_{zz}\varepsilon_{zz})\,dV,
\end{aligned}
$$

and therefore the strain energy density referred to principal axes is

$$
W = \tfrac{1}{2}(\tau_{xx}\varepsilon_{xx} + \tau_{yy}\varepsilon_{yy} + \tau_{zz}\varepsilon_{zz}).
$$

In the general case of arbitrary (in general, nonprincipal) axes, this expression assumes the form

$$
\begin{aligned}
W &= \tfrac{1}{2}(\tau_{xx}\varepsilon_{xx} + \tau_{xy}\varepsilon_{xy} + \tau_{xz}\varepsilon_{xz}) \\
&\quad + \tfrac{1}{2}(\tau_{yx}\varepsilon_{yx} + \tau_{yy}\varepsilon_{yy} + \tau_{yz}\varepsilon_{yz}) \\
&\quad + \tfrac{1}{2}(\tau_{zx}\varepsilon_{zx} + \tau_{zy}\varepsilon_{zy} + \tau_{zz}\varepsilon_{zz}),
\end{aligned}
$$

or, in abbreviated notation,

$$
W = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\tau_{ij}\varepsilon_{ij}. \tag{27}
$$

In view of the relations in Eqs. (24) and (27), the expression for strain energy density can be written in the form

$$
W = \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3}C_{ijkl}\varepsilon_{ij}\varepsilon_{kl}. \tag{28a}
$$

In the case of an isotropic elastic material [see Eq. (26)], this equation reduces to

$$
W = \frac{1}{2}\left[\lambda(\varepsilon_{11}+\varepsilon_{22}+\varepsilon_{33})^2 + 2G\sum_{i=1}^{3}\sum_{j=1}^{3}\varepsilon_{ij}\varepsilon_{ij}\right], \tag{28b}
$$

where

$$
\lambda = E/(1+v)(1-2v).
$$

Thus, with reference to Eq. (28), we note that the strain energy density is a quadratic function of the strain tensor components, and W vanishes when the strain field vanishes. Equation (28) serves as a potential (generating) function for the generation of the stress field, that is,

$$
\tau_{ij} = \frac{\partial W(\varepsilon_{ij})}{\partial \varepsilon_{ij}} = \sum_{k=1}^{3}\sum_{l=1}^{3}C_{ijkl}\varepsilon_{kl}; \quad i = 1, 2, 3; \quad j = 1, 2, 3 \tag{29}
$$

[see Eq. (24)]. The concept of strain energy serves as the starting point for many useful and important investigations in elasticity theory and its applications. For details, the reader is referred to the extensive literature, a small selection of which can be found in the Bibliography.

VI. EQUILIBRIUM AND THE FORMULATION OF BOUNDARY VALUE PROBLEMS

External agencies usually deform a solid by two distinct types of loadings: (a) surface tractions and (b) body forces. Surface tractions act by virtue of the application of normal and shearing stresses to the surface of the solid, while body forces act upon the interior, distributed mass of the solid. For example, a box resting on a table is subjected to (normal) surface traction forces at the interface between tabletop and box bottom, whereas gravity causes forces to be exerted upon the contents of the box.

Consider a solid body B bounded by the surface S in a state of static equilibrium. Then at every internal point of B, these partial differential equations must be satisfied:

$$
\begin{aligned}
\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} + F_x &= 0 \\
\frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} + F_y &= 0 \\
\frac{\partial \tau_{zx}}{\partial x} + \frac{\partial \tau_{zy}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z} + F_z &= 0,
\end{aligned} \tag{30}
$$

where $\tau_{xy} = \tau_{yx}$, $\tau_{yz} = \tau_{zy}$, $\tau_{zx} = \tau_{xz}$, and $\mathbf{F} = F_x\mathbf{e}_x + F_y\mathbf{e}_y + F_z\mathbf{e}_z$ is the body force vector per unit volume.

The admissible boundary conditions associated with Eq. (30) may be stated in the form:

$$
\mathbf{T} \equiv (T_1, T_2, T_3) \qquad \text{on} \qquad S_1
$$

and

$$
\mathbf{u} \equiv (u, v, w) \qquad \text{on} \qquad S_2, \tag{31}
$$

where $\mathbf{T}$ is the surface traction vector [see Eq. (4)], $\mathbf{u}$ is the displacement vector, and $S = S_1 + S_2$ denotes the bounding surface of the solid.

The solution of a problem in (three-dimensional) elasticity theory requires the determination of

$$
\left.\begin{array}{l}\text{the displacement vector field } \mathbf{u} \\ \text{the stress tensor field } \tau_{ij} \\ \text{and the strain tensor field } \varepsilon_{ij}\end{array}\right\} \quad \text{in } B. \tag{32}
$$

This solution is required to satisfy the equations of equilibrium [Eq. (30)], the equations of compatibility [Eq. (23)], the strain-displacement relations [Eq. (20)], and the stress–strain relations [Eq. (26) or (24)], as well as the boundary conditions [Eq. (31)]. This is a formidable task, even for relatively simple geometries and boundary conditions, and the exact or approximate solution requires extensive use of advanced analytical as well as numerical mathematical methods in most cases.

VII. EXAMPLES

A. Example A

We consider an elastic cylinder of length L with an arbitrary cross section. The cylinder is composed of a linearly elastic, isotropic material with Young's modulus E and Poisson's ratio ν. The cylinder is inserted into a perfectly fitting cavity in a rigid medium, as shown in Fig. 6, and subjected to a uniformly distributed normal stress $\tau_{zz} = T$ on the free surface at z = L. We assume that the bottom of the cylinder remains in smooth contact with the rigid medium, and that the lateral surfaces between the cylinder and the rigid medium are smooth, thus capable of transmitting normal surface tractions only. Moreover, normal displacements over the lateral surfaces are prevented. Thus, we have the displacement field

$$u = v = 0, \qquad w = (\delta/L)z,$$

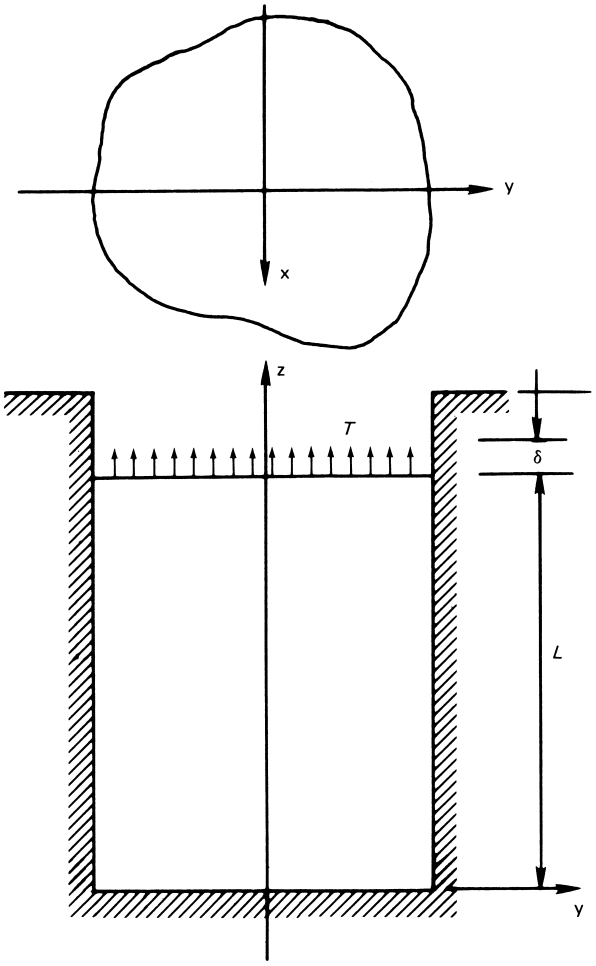


FIGURE 6 Transversely constrained cylinder.

where δ is the z displacement of the top of the cylinder. With the aid of Eq. (20), we obtain the strain field

$$\begin{aligned} \varepsilon_{xx} = \varepsilon_{yy} = 0; \qquad & \varepsilon_{zz} = \delta/L; \\ \varepsilon_{ij} \equiv 0, \qquad & i \neq j. \end{aligned} \tag{33}$$

In view of Eqs. (26) and (33), we have

$$\tau_{xx} - \nu(\tau_{yy} + \tau_{zz}) = 0,$$
$$\tau_{yy} - \nu(\tau_{xx} + \tau_{zz}) = 0,$$

and

$$\tau_{zz} - \nu(\tau_{yy} + \tau_{xx}) = E(\delta/L),$$

and therefore,

$$\begin{aligned} \tau_{xx} = \tau_{yy} &= \frac{\nu}{1-\nu}\tau_{zz} \\ \tau_{zz} &= E\frac{\delta}{L}\frac{(1-\nu)}{(1-2\nu)(1+\nu)} = T, \\ \tau_{ij} &= 0 \qquad \text{for} \quad i \neq j. \end{aligned} \tag{34}$$

In the case of a copper cylinder, we have (see Table I) $\nu = 0.34$, $E = 8.96 \times 10^{10}$ Pa; and for an axial strain $\varepsilon_{zz} = \delta/L = 0.0005$, we readily obtain

$$\tau_{xx} = \tau_{yy} = 35.53 \times 10^6 \text{ Pa}$$

and

$$\tau_{zz} = 68.9 \times 10^6 \text{ Pa}.$$

Thus, when we compress the copper cylinder with a stress $\tau_{zz} = T = -68.9 \times 10^6$ Pa, there will be induced a lateral compressive stress $\tau_{xx} = \tau_{yy} = -35.53 \times 10^6$ Pa. We note that the strain field [Eq. (33)] satisfies the equations of compatibility [Eq. (23)] and the stress field [Eq. (34)] satisfies the equations of equilibrium [Eq. (30)] provided the body force vector field **F** vanishes (or is negligible).

B. Example B

We consider the case of plane, elastic pure bending (or flexure) of a beam by end couples as shown in Fig. 7. In the reference state, the z axis and the beam longitudinal axis coincide. The cross section of the beam (normal to the z axis) is constant and symmetrical with respect to the y axis. Its area is denoted by the symbol A, and the centroid of A is at $(0, 0, z)$. The beam is acted upon by end moments $M_x = M$ such that

$$M_x = \int_A \tau_{zz} y \, dA = M$$

and

$$M_y = \int_A \tau_{zz} x \, dA = 0.$$

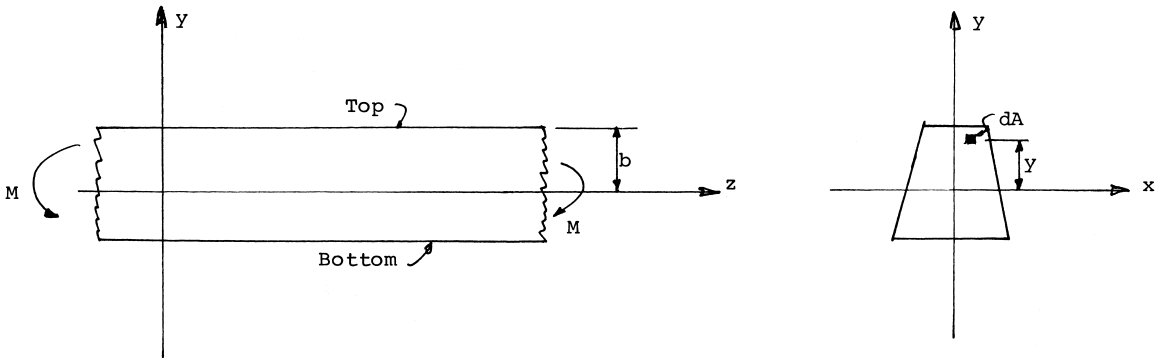


FIGURE 7 Pure bending of a beam.

The present situation suggests the stress field

$$\begin{bmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \tau_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \tau_{zz} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \dfrac{My}{I} \end{bmatrix}, \tag{35}$$

where $I = \int_A y^2\, dA$, on account of physical reasoning and (elementary) Euler-Bernoulli beam theory. Upon substitution of Eq. (35) into Eq. (26), and in view of Eq. (20), we obtain

$$\begin{aligned} \varepsilon_{xx} &= -\frac{\nu}{E}\tau_{zz} = -\frac{\nu}{E}\frac{M}{I}y = \frac{\partial u}{\partial x} \\ \varepsilon_{yy} &= -\frac{\nu}{E}\tau_{zz} = -\frac{\nu}{E}\frac{M}{I}y = \frac{\partial v}{\partial y} \\ \varepsilon_{zz} &= \frac{\tau_{zz}}{E} = \frac{M}{EI}y = \frac{\partial w}{\partial z}, \end{aligned} \tag{36}$$

and all shearing strains vanish.

We now integrate the partial differential equations in (36), subject to the following boundary conditions: At $(x, y, z) = (0, 0, 0)$ we require $u = v = w = 0$ and

$$\frac{\partial u}{\partial z} = \frac{\partial y}{\partial z} = \frac{\partial u}{\partial y} = 0.$$

Thus, the beam displacement field is given by

$$\begin{aligned} u &= -\frac{M\nu}{EI}xy \\ v &= -\frac{M}{2EI}[z^2 + \nu(y^2 - x^2)] \\ w &= \frac{M}{EI}yz. \end{aligned} \tag{37}$$

We note that the strain field (36) satisfies the equation of compatibility (23) and the stress field (35) satisfies the equations of equilibrium (30) provided the body force vector field $\mathbf{F}$ vanishes (or is negligible).

With reference to Fig. 7, in the reference configuration, the top surface of the beam is characterized by the plane $y = b$. Subsequent to deformation, the top surface of the beam is characterized by

$$v = -\frac{M}{2EI}(z^2 - \nu x^2) - \frac{\nu M b^2}{2EI}, \tag{38}$$

and for $(x, y, z) = (0, b, 0)$ we have

$$v(0, b, 0) = -\frac{\nu M b^2}{2EI}.$$

We now write Eq. (38) in the form

$$V = v + \frac{\nu M b^2}{2EI} = -\frac{M}{2EI}(z^2 - \nu x^2), \tag{39}$$

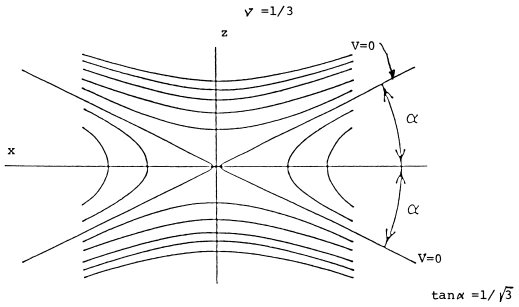
and we note that V denotes the deflection of the (originally) plane top surface of the beam. The contour lines $V =$ constant of this saddle surface are shown in Fig. 8a. We note that the contour lines consist of two families of hyperbolas, each having two branches. The asymptotes are straight lines characterized by $V = 0$, so that $\tan \alpha = z/x = \sqrt{\nu}$.

An experimental technique called holographic interferometry is uniquely suited to measure sufficiently small deformations of a beam loaded as shown in Fig. 7. In Fig. 8b we show a double-exposure hologram of the deformed top surface of a beam loaded as shown in Fig. 7. This hologram was obtained by the application of a two (light) beam technique, utilizing Kodak Holographic 120-02 plates. The laser was a 10-mW He-Ne laser, 632.8 nm, with beam ratio 4:1. The fringe lines in Fig. 8b correspond to the contour lines of Fig. 8a. The close correspondence between theory and experiment is readily observed. We also note that this technique results in the nondestructive, experimental determination of Poisson's ratio ν of the beam.

C. Example C

We wish to find the displacement, stress field, and strain field in a spherical shell of thickness $(b - a) > 0$ subjected to uniform, internal fluid (or gas) pressure p. The shell is

(a)



(b)

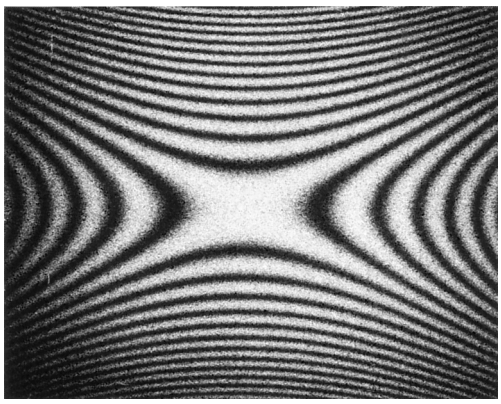


FIGURE 8 (a) Contour lines; V, constant. (b) Double-exposure hologram of deformed plate surface. (Holographic work was performed by P. Malyak in the laboratory of D. P. Malone, Department of Electrical Engineering, State University of New York at Buffalo.) [This hologram is taken from Reismann, H., and Pawlik, P. S. (1980). "Elasticity: Theory and Applications," Wiley (Interscience), New York.]

bounded by concentric spherical surfaces with outer radius $r = b$ and inner radius $r = a$, and we designate the center of the shell by O. In view of the resulting point-symmetric displacement field, there will be no shear stresses acting upon planes passing through O and upon spherical surfaces $a \le r \le b$. Consequently, at each point of the shell interior, the principal stresses are radial tension (or compression) τ_{rr} and circumferential tension (or compression) $\tau_{\theta\theta}$, the latter having equal magnitude in all circumferential directions.

To obtain the pertinent equation of equilibrium, we consider a volume element (free body) bounded by two pairs of radial planes passing through O, each pair subtending a (small) angle $\Delta\theta$, and two spherical surfaces with radii r and $r + \Delta r$. Invoking the condition of (radial) static equilibrium, we obtain

$$(\tau_{rr} + \Delta\tau_{rr})[(r + \Delta r)\Delta\theta]^2 - \tau_{rr}(r\Delta\theta)^2 = 2\tau_{\theta\theta}\left(r + \frac{\Delta r}{2}\right)\Delta r(\Delta\theta)^2.$$

We now divide this equation by $(\Delta\theta)^2$ and Δr then take the limit as $\Delta r \to 0$ and $\Delta\tau_{rr}/\Delta_r \to d\tau_{rr}/dr$. The result of these manipulations is the stress equation of equilibrium

$$\frac{d\tau_{rr}}{dr} + \frac{2}{r}(\tau_{rr} - \tau_{\theta\theta}) = 0. \tag{40}$$

In view of the definition of strain in Section III, the strain-displacement relations for the present problem are

$$\varepsilon_{rr} = \frac{(dr + du) - dr}{dr} = \frac{du}{dr}$$
$$\varepsilon_{\theta\theta} = \frac{2\pi(r + u) - 2\pi r}{2\pi r} = \frac{u}{r}, \tag{41}$$

where the letter u denotes radial displacement. For our present purpose, we now write Hooke's law (26) in the following form:

$$\tau_{rr} = (\lambda + 2G)\varepsilon_{rr} + 2\lambda\varepsilon_{\theta\theta}$$
$$\tau_{\theta\theta} = 2(\lambda + G)\varepsilon_{\theta\theta} + \lambda\varepsilon_{rr}, \tag{42}$$

where

$$\lambda = \frac{E\nu}{(1 + \nu)(1 - 2\nu)} = \frac{2G\nu}{(1 - 2\nu)}.$$

If we substitute Eq. (41) into Eq. (42) and then substitute the resulting equations into Eq. (40), we obtain the displacement equation of equilibrium

$$\frac{d^2u}{dr^2} + \frac{2}{r}\frac{du}{dr} - \frac{2}{r^2}u = 0. \tag{43}$$

The spherical shell has a free boundary at $r = b$ and is stressed by internal gas (or liquid) pressure acting upon the spherical surface $r = a$. Consequently, the boundary conditions are

$$\tau_{rr}(a) = -p, \tag{44}$$

where $p \ge 0$ and $\tau_{rr}(b) = 0$. The solution of the differential equation (43) subject to the boundary conditions (44) is

$$u = \frac{pa^3r}{3K(b^3 - a^3)} + \frac{pa^3b^3}{4G(b^3 - a^3)r^2}, \qquad a \le r \le b, \tag{45}$$

where $K = E/[3(1 - 2\nu)] = (3\lambda + 2G)/3$ is the modulus of volume expansion, or bulk modulus. Upon substitution of Eq. (45) into Eq. (41), we obtain the strain field

$$\varepsilon_{rr} = \frac{pa^3}{3K(b^3 - a^3)} - \frac{pa^3b^3}{2G(b^3 - a^3)r^3}$$
$$\varepsilon_{\theta\theta} = \frac{pa^3}{3K(b^3 - a^3)} + \frac{pa^3b^3}{4G(b^3 - a^3)r^3}, \tag{46}$$

and upon substitution of Eq. (46) into Eq. (42), we obtain the stress field

$$\tau_{rr} = \frac{pa^3}{(b^3 - a^3)}\left[1 - \left(\frac{b}{r}\right)^3\right] = \sigma_2 = \sigma_3 \leq 0$$
$$\tau_{\theta\theta} = \frac{pa^3}{(b^3 - a^3)}\left[1 + \frac{1}{2}\left(\frac{b}{r}\right)^3\right] = \sigma_1 \geq 0. \tag{47}$$

We also note the following relations:

$$\tau_{rr} + 2\tau_{\theta\theta} = \frac{3pa^3}{(b^3 - a^3)}, \quad \varepsilon_{rr} + 2\varepsilon_{\theta\theta} = \frac{pa^3}{K(b^3 - a^3)},$$
$$\frac{\tau_{rr} + 2\tau_{\theta\theta}}{\varepsilon_{rr} + 2\varepsilon_{\theta\theta}} = 3K. \tag{48}$$

With reference to Eq. (16), the octahedral shear stress is

$$\tau_0 = \frac{1}{3}\left[(\sigma_1 - \sigma)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2\right]^{1/2}$$
$$= \frac{\sqrt{2}}{2}\frac{pa^3}{(b^3 - a^3)}\left(\frac{b}{r}\right)^3, \tag{49}$$

and the maximum shear stress (as a function of r) is

$$\tau_{\max} = \frac{1}{2}(\sigma_1 - \sigma_3) = \frac{3}{4}\frac{pa^3}{(b^3 - a^3)}\left(\frac{b}{r}\right)^3, \tag{50}$$

and we note that for the present case we have $\tau_0/\tau_{\max} = (2\sqrt{2})/3 \cong 0.9428$ and [see Eq. (19)]

$$1 < \sqrt{\frac{3}{2}}\frac{\tau_0}{\tau_{\max}} = \frac{2}{\sqrt{3}}. \tag{51}$$

We now apply the failure criterion due to Hencky-Mises (see Section IV): Yielding will occur when $3\tau_0 = \sqrt{2}Y$, where Y denotes the yield stress in simple tension of the shell material. Upon application of this criterion and with the aid of Eq. (49), we obtain

$$p = \frac{2}{3}\frac{(b^3 - a^3)}{a^3}\left(\frac{r}{b}\right)^3 Y, \tag{52}$$

and the smallest value of p results when $r = a$. Thus we conclude that the Hencky-Mises failure criterion predicts yielding on the surface $r = a$ when

$$p = \frac{2}{3}\left[1 - \left(\frac{a}{b}\right)^3\right]Y. \tag{53}$$

The criterion due to Tresca (see Section IV) predicts failure when $\tau_{\max} = Y/2$. With the aid of Eq. (50), this results again in Eq. (53), and we conclude that for the present example, the failure criteria of Hencky-Mises and Tresca predict the same pressure at incipient failure of the shell given by the formula (53).

SEE ALSO THE FOLLOWING ARTICLES

ELASTICITY, RUBBERLIKE • FRACTURE AND FATIGUE • MECHANICS, CLASSICAL • MECHANICS OF STRUCTURES • NUMERICAL ANALYSIS • STRUCTURAL ANALYSIS, AEROSPACE

BIBLIOGRAPHY

Boresi, A. P., and Chong, K. P. (1987). "Elasticity in Engineering Mechanics," Elsevier, Amsterdam.

Brekhovskikh, L., and Goncharov, V. (1985). "Mechanics of Continua and Wave Dynamics," Springer-Verlag, Berlin and New York.

Filonenko-Borodich, M. (1963). "Theory of Elasticity," Peace Publishers, Moscow.

Fung, Y. C. "Foundations of Solid Mechanics," Prentice-Hall, Englewood Cliffs, NJ.

Green, A. E., and Zerna, W. (1968). "Theoretical Elasticity," 2nd ed., Oxford Univ. Press, London and New York.

Landau, L. D., and Lifshitz, F. M. (1970). "Theory of Elasticity" (Vol. 7 of Course of Theoretical Physics), 2nd ed., Pergamon, Oxford.

Leipholz, H. (1974). "Theory of Elasticity," Noordhoff-International Publications, Leyden, The Netherlands.

Lur'e, A. I. (1964). "Three-Dimensional Problems of the Theory of Elasticity," Wiley (Interscience), New York.

Novozhilov, V. V. (1961). "Theory of Elasticity," Office of Technical Services, U.S. Department of Commerce, Washington, D.C.

Parkus, H. (1968). "Thermoelasticity," Ginn (Blaisdell), Boston.

Parton, V. Z., and Perlin, P. I. (1984). "Mathematical Methods of the Theory of Elasticity," Vols. I and II, Mir Moscow.

Reismann, H., and Pawlik, P. S. (1974). "Elastokinetics," West, St. Paul, Minn.

Reismann, H., and Pawlik, P. S. (1980). "Elasticity: Theory and Applications," Wiley (Interscience), New York.

Solomon, L. (1968). "Elasticité Linéaire," Masson, Paris.

Southwell, R. V. (1969). "An Introduction to the Theory of Elasticity," Dover, New York.

Timoshenko, S. P., and Goodier, J. M. (1970). "Theory of Elasticity," 3rd ed., McGraw-Hill, New York.



Electromagnetic Compatibility

J. F. Dawson
A. C. Marvin
C. A. Marshman
University of York

I. Sources of Electromagnetic Interference
II. Effects of Interference
III. Interference Coupling Paths and Their Control
IV. Design for Electromagnetic Compatibility
V. Electromagnetic Compatibility Regulations and Standards
VI. Measurement and Instrumentation

GLOSSARY

Antenna factor The factor by which the received voltage at a specified load is multiplied to determine the received field at the antenna.
Common-mode current/voltage The component of current/voltage which exists equally and in the same direction on a pair of conductors or multiconductor bundle, i.e., the return is via a common ground connection (cf. differential mode).
Crosstalk Unintentional transfer of energy from one circuit to another by inductive or capacitive coupling or by means of a common impedance (e.g., in a common return conductor).
Differential mode current/voltage The component of current/voltage which exists equally and in opposite directions on a pair of conductors (cf. common mode).
Shielding effectiveness The ratio of electric or magnetic field strength without a shield to that with the shield present (larger numbers mean better shielding).
Skin depth The depth of the layer in which radiofrequency current flows on the surface of a conductor.
Skin effect The confinement, at high frequencies, of current to a thin layer close to the surface of a conductor.
Source The source of electromagnetic interference.
Victim A circuit or system affected by electromagnetic interference.

ELECTROMAGNETIC COMPATIBILITY (EMC) is the ability of electrical and electronic systems to coexist with each other without causing or suffering from malfunction due to electromagnetic interference (EMI) from each other or from natural causes. As we rely more and more upon electronic systems for the day-to-day operation of our factories, houses, and transport systems, the need to achieve electromagnetic compatibility has increased in importance. This has resulted in the design, analysis, and measurement techniques discussed in this article.

The limits of electromagnetic (EM) emissions from equipment and immunity to EMI that an equipment must tolerate in an operating environment are determined by standards organizations, in particular, the International Electrotechnical Commission (IEC) and its CISPR committee (Comité International Special Perturbations Radioelectrique). The guidelines laid down in the standards may be enforced through regulations.

I. SOURCES OF ELECTROMAGNETIC INTERFERENCE

A. Natural Sources

1. Electrostatic Discharge

When differing materials are in sliding contact one material may lose electrons to the other—this is the *triboelectric effect*. This results in a buildup of electrical charge. The electric field due to the charge can cause electrical breakdown of the air (or other insulating material) surrounding the source of the charge, resulting in an electrostatic discharge (ESD).

The rate of charge transfer depends on the materials in contact. Electrostatic discharge can be reduced by using materials which are closely matched in the triboelectric series or by using materials with a low conductivity which allow the charge to leak away before it accumulates sufficiently to discharge due to a breakdown of insulation.

A common cause of electrostatic discharge is the use of synthetic clothing and furniture. Electric charge is induced on the human body due to friction between clothing or shoes and furniture or floor coverings; the body capacitance (a few hundred picofarads) can charge to voltages as high as 15 kV. When the body comes in close proximity to electronic equipment a spark between the body and metal on the equipment may occur. This can result in a large current flow with a very fast rise time (<1 nsec) and a duration of about 100 nsec, which may disrupt or damage the electronic equipment as well as radiating electromagnetic energy which may disturb nearby equipment. The fast-rise-time, initial peak is due to the discharge of the finger and arm, while the slower, secondary peak is due to the discharge of the remainder of the body (Fig. 1).

Electrostatic discharge can also occur on aircraft due to air friction and on satellites due to direct bombardment with charged particles.

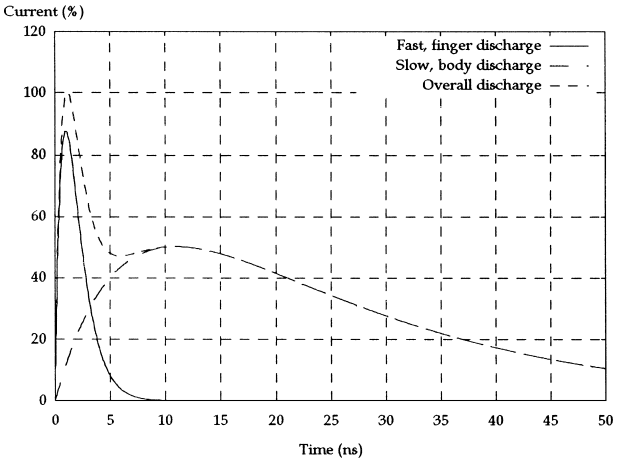


FIGURE 1 Approximate current waveform for electrostatic discharge.

2. Lightning

Lightning is the result of the ionization of the air due to charge accumulated in clouds. This is thought to be due to a triboelectric effect between ice crystals. Lightning has a much larger energy than the electrostatic discharge phenomenon described above.

Lightning discharges have a rise time of the order of 1 μsec and decay in about 50 μsec. Lightning strikes can induce large currents (up to 100 kA) and voltages (up to 100 kV) in conductors and may therefore be a source of electromagnetic interference to electronic systems.

3. Solar Storms

Solar storms can induce large currents in power networks. This low-frequency interference phenomenon resulted in a blackout of the Hydro-Quebec power grid in Canada during the 1989 solar maximum; it has been of much interest to power companies worldwide, who have spent considerable resources to in harden their distribution systems to prevent similar occurrences during the 2000 solar maximum.

B. Man-Made Sources

1. Intentional Sources

Radio and radar transmitters, industrial, scientific, and medical (ISM) equipment using high-power radiofrequency energy, and microwave ovens and other equipment which produces significant radiofrequency fields intentionally are also sources of interference for systems which may be susceptible to their emissions.

The proliferation of mobile phones is a significant source of interference. Their use must be controlled near sensitive systems such as medical monitoring equipment and in aircraft (to prevent interference with navigation aids). The increase in wireless networking for portable computing devices is likely to increase this problem.

2. Unintentional Sources

Most electrical and electronic equipment has the potential to cause electromagnetic interference. Particular sources include:

- Electrical circuit breakers, contactors, and relays, which are likely to draw arcs as they open or close
- Arcs such as those created by welding equipment and furnaces and discharges such as those in fluorescent lighting
- Brushgear in electrical machinery
- Solid state switching circuits ranging from logic circuits to switched-mode power converters
- Radiofrequency oscillators (including unintentional ones).

II. EFFECTS OF INTERFERENCE

A. Interference with Radio Communications and Navigation Aids

EMC began with the need to prevent electrical noise generated by trams (trolley cars) and automobile ignition systems from interfering with broadcast radio transmission. The main factor driving limits on electromagnetic emissions in EMC standards is still the prevention of radio interference. The interference from mobile phones and portable electronic equipment has become a problem in certain sensitive environments. In aircraft, where sensitive radio receivers are used for communications and navigation, the use of portable electronic equipment is prohibited during critical phases of flight (i.e., takeoff and landing).

B. Malfunction of Electronic Systems

Equipment which does not interfere with radio reception is unlikely to interfere with other electronic systems; however, many of the sources of interference described above produce a large enough disturbance to interfere with the normal operation of electronic systems.

1. Demodulation and Intermodulation

Radiofrequency interference is often outside the passband of a circuit and does not directly interfere with the wanted signal. However, all active components have a degree of nonlinearity, which means that radiofrequency interference which enters a circuit can be demodulated to produce signals within the passband of the circuit. This is the most common cause of interference effects in analog circuits.

If more than one interfering frequency is present, sum and difference frequencies of the fundamental components and their harmonics (intermodulation products) are generated by any nonlinearities, which can result in new frequency components that may be within the passband of the circuit.

2. Data Corruption

The presence of interference in digital circuits can induce timing jitter (may cause failure due to violation of timing constraints) and eventual direct corruption of data in digital circuits (noise margin).

3. Damage

High levels of interference can cause damage to components which may result in their failure or reduced reliability (latent failures). Electrostatic discharge is one of the most common causes of damage to electronic components both in their handling (prior to manufacture) and in service. The high-intensity radiated fields (HIRF) in the vicinity of radio, television, and radar transmitters can induce sufficient energy in electronic systems to cause damage; this is of particular concern to the aerospace industry, where sensitive electronic systems must operate in the vicinity of high-power radars and radio transmitters.

III. INTERFERENCE COUPLING PATHS AND THEIR CONTROL

A. Conducted Interference

At frequencies below 30 MHz interference can propagate efficiently along power circuits within buildings and other installations. Above 30 MHz the attenuation in electrical wiring limits the propagation of interference and direct radiation often provides a lower loss path.

Conducted interference can be resolved into differential and common-mode components. Figure 2 shows two grounded enclosures linked by two wires. A source in one enclosure drives a load in the other and, as expected, a current I_{dm} (the *differential mode* current) flows in each wire in opposing directions. No current flows in the ground. Figure 3 shows a circuit in which a *common-mode* current

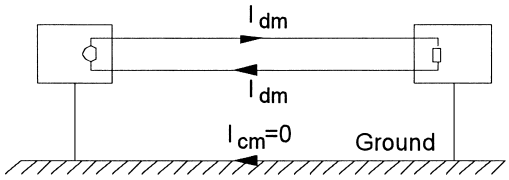


FIGURE 2 The flow of differential mode current in a two-wire connection between grounded enclosures.

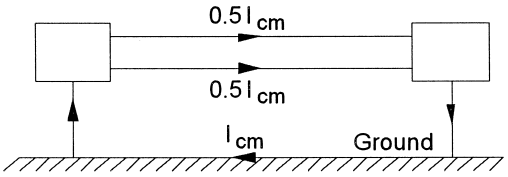


FIGURE 3 The flow of common-mode current in a two-wire connection between grounded enclosures.

flows equally, in the same direction, in each of two conductors, returning through the common ground connection. The flow of common mode current is often caused by external fields inducing current in the ground loop or due to the two wires having different impedances to ground, resulting in the generation of an EMF which can drive a current around the ground loop. Potential differences between different grounding points due to earth-leakage currents and lightning strikes, can also be a source of common mode current. In practice, both differential and common-mode currents are present in most cases (Fig. 4).

1. Sources

Switch mode power supplies, inverters, and speed controllers generate switching transients, which may appear as conducted interference in other systems. Linear power supplies generate harmonic currents, which increase losses in the distribution system but do not directly affect other electronic systems.

Electrical switchgear is a source of conducted interference due to the arcing which occurs when contacts are operated. The rapid establishment and breaking of an arc (showering arc) which can occur when switching loads with inductive and capacitive elements can result in a burst of short (5 nsec rise time, 50 nsec width), high-voltage (several kilovolts) transients, lasting for a few tens of milliseconds, known as a fast transient burst. Damped oscillatory transients at frequencies of 100 kHz to 1 MHz can also be generated during contact operation.

Brushgear on electrical machines can cause broadband conducted electrical noise.

Induced currents in ground, power supply, and signal wiring due to lightning and electrostatic discharge can cause significant interference and damage to electronic systems.

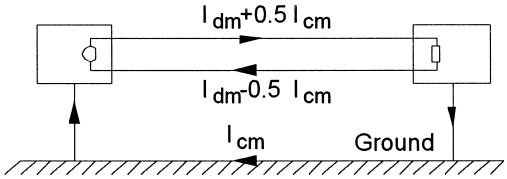


FIGURE 4 The total current: the sum of common-mode and differential mode currents.

2. Control

Filters, transient suppressors, and isolation techniques may be used to reduce the amplitude of conducted interference at both the source and victim. Shielded cables may help to reduce coupling of interference between cables running in close proximity (e.g., in a cable duct).

It is not practical to completely filter transients from electrical switchgear, so potential victim equipment must have inherent immunity. Care must be taken to ensure that interference does not bypass the protection circuits (e.g., by direct coupling between the input and output connections).

Filters work by using frequency-dependent impedances (e.g., capacitors and inductors) consisting of series elements, which are intended to reduce the flow of interfering current, and shunt elements, which are intended to allow the interfering current to bypass the *victim* circuit. Clearly a filter can only be effective when the spectrum of the interference differs from the spectrum of the desired signal or power source. The large transient voltages that can occur at the input of filters used in EMC applications mean that care must be take to ensure that inductor cores do not saturate (reducing their effectiveness) and that the dielectric strength of capacitors is not exceeded.

Safety must be considered for filters used on power circuits. In particular the presence of capacitors between line and chassis ground can cause an equipment chassis to become “live” if the chassis ground becomes disconnected. Capacitors to be used in line-power circuits are subject to regulatory control which limits the size of the capacitor to limit the current flow in case of electric shock via a “live” chassis due to a disconnected ground. These capacitors must be self-healing so as to correct any dielectric damage due to overvoltage transients.

Filters for EMC differ from filters for communications and signal processing because they operate in a less well controlled environment—source and load impedances may vary rapidly with frequency (a few ohms to a few kilo-ohms and any phase angle is typical). In order to control the filter behavior with a wide range of load and source impedances, lossy elements are often incorporated into high-quality filters. These include lossy ferrite cores and simple resistors.

Transient suppressors are used in conjunction with filters to minimize the effect of large-amplitude transients on electronic equipment. The spark gap is used widely as a transient suppressor and has the advantages of low capacitance, high impedance when not activated, and the ability to shunt very high currents when the arc is struck;

however, it is relatively slow in operation (on the order of microseconds) and metal vapor produced from the electrodes when an arc is struck is deposited in the envelope and leads to a falling resistance and eventual failure. The low voltage required to sustain an arc means that some means of quenching the arc must be used in power circuits (e.g., a fuse or contact breaker). Metal oxide varistors have a faster response than a spark gap, and their inherent capacitive nature can be an advantage in some applications. The varistor has a constant voltage characteristic and may be used in power circuits to limit transients without any quenching mechanism. The voltage and current capabilities of a varistor are lower than those of a spark gap. Avalanche diodes optimized for speed are used as transient suppressors to protect sensitive solid state circuitry; these are essentially low-voltage devices (a few tens of volts) with a very fast switching time (nanoseconds). In practical suppression systems all three devices may be used in conjunction with filter elements to prevent damage from high-energy transients such as those induced by nearby lightning strikes.

In the case of common-mode currents, both filters and transient suppressors rely on bypassing some of the interfering signal to a ground or chassis connection, rather than having it flow through the internal ground circuitry. A good connection to a metal chassis for transient suppressors and filters is essential for their correct operation.

Signal circuits may be protected from common-mode, conducted interference by means of opto- or transformer-based isolators, which break the ground loop and prevent the flow of common-mode current.

Shielded cables are useful for minimizing the effect of the coupling of common-mode, conducted interference between cables in close proximity. Shielding is also appropriate when the spectra of the signal or power connection overlap the interference spectrum, so that filters are not applicable. The common-mode currents flow on the cable shield, rather than the signal wires, reducing the effect of interference. However, the flow of large interference currents on cable shields can cause problems in itself. If a cable shield is used as a zero-volt reference for the signal wires, then any potential difference along the cable shield due to the flow of common-mode current appears in series with the signal voltages. Also, the presence of a large circulating current in the cable shield can cause safety problems. This should be addressed by the proper safety bonding of equipment and building grounding systems.

B. Radiated Interference

1. Emissions from Electronic Equipment

The electromagnetic radiation from an electrical circuit increases with the rate of change of current and/or voltage in the circuit. Efficient antennas must be a significant fraction of a wavelength (e.g., half- or quarter-wave resonance), so that equipment begins to radiate efficiently when it is of the order of one half-wavelength large or has cables attached of that length. Typical desktop equipment has leads of the order of 1 m attached and so begins to radiate electromagnetic noise efficiently in the VHF band (Fig. 5). Tracks on printed circuit boards (PCBs), heatsinks, apertures, and seams (joints) in equipment cases and other small structures can also become efficient antennas when their length becomes a significant fraction of a wavelength. In desktop equipment this radiation mechanism becomes significant in the UHF band. With microprocessor operating speeds moving into the gigahertz region, radiation from small structures is becoming a significant factor.

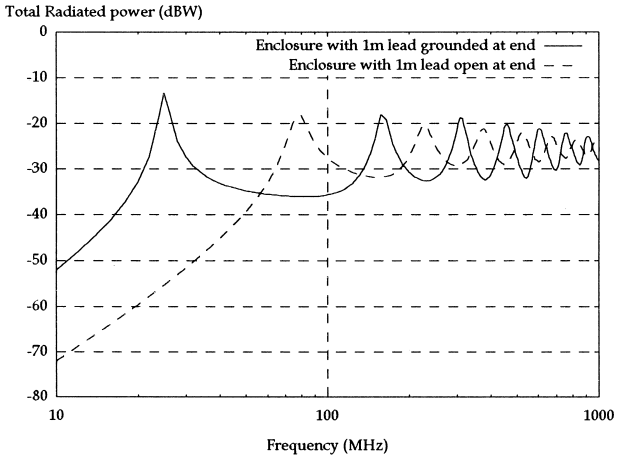


FIGURE 5 Total radiated power from an equipment enclosure with a 1-m lead excited by a 1-V (common-mode) source. Note the drop in resonant frequency when the lead is grounded due to image currents in the ground.

2. Susceptibility of Electronic Equipment

Radiated interference enters electronic equipment through cables, apertures, seams, etc.—the same paths that allow emissions. Although the propagation mechanisms are identical, the circuits that are likely to be affected by interference entering a system are often not the same as those likely to cause radiated emissions. Therefore measures taken to suppress emissions do not necessarily have any effect on the susceptibility of equipment to external sources of interference and vice versa.

3. Control

Radiated interference is controlled in part by means of filters, shielding, and the physical layout of a system. The careful design of software and circuits is also an important factor.

Filters serve to prevent unwanted frequency signals passing between a system and external cabling which may act as an antenna. A significant example is the use of a ferrite bead on cabling to reduce the common-mode currents on the bundle or screen. Uncontrolled common-mode currents on cables are the most common cause of radiation from equipment in the VHF band. The common-mode current may be induced by imbalance in the signal and return connections in the cable, potential differences in the grounding structure of equipment, or currents coupled via internal cable looms. A cable that is an efficient radiator is also an effective receiver of interference—the path into or out of a piece of equipment is reciprocal—so a filter will have the same effect on immunity as emissions.

Screened cables can greatly reduce cable radiation and ingress of interference via cables. At radio frequencies the currents induced by internal wires tend to flow on the inside of the screen, while currents induced by external fields flow on the outside due to the skin effect; the two tend not to interact. The imperfections in braided screens limit their effectiveness compared with a solid screen as frequency increases. The performance of a cable screen is specified by its transfer impedance Z_t such that the equivalent voltage source V_s on the inner conductor due to a current I_s on the outside of the screen for a an element of cable of length δl is given by

$$V_s = I_s Z_t \delta l.$$

Figure 6 shows the typical variation of Z_t with frequency for solid and braided shielded cables.

In many cases the limiting factor in the performance of a screened cable is the manner in which its screen is connected to the equipment enclosure. A good-quality connector which maintains a 360-deg connection of the screen

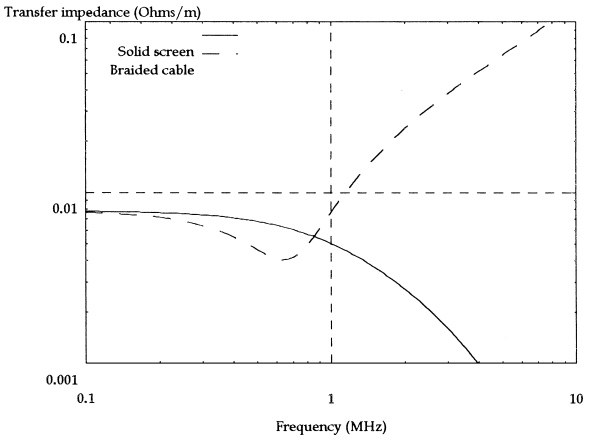


FIGURE 6 Transfer impedance of typical solid and good braided screened cables.

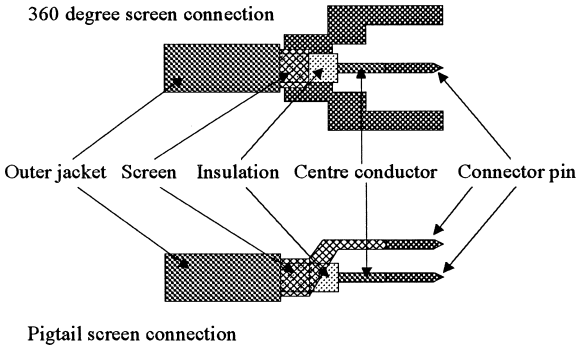


FIGURE 7 A connector with 360-deg cable termination and a connector with a pigtail screen connection.

to the equipment enclosure is required to realize the full performance of a cable. A termination in which the screen is gathered into a loop (often known as a pigtail) before connection to the enclosure will significantly degrade the performance of the screen (Fig. 7).

A conductive enclosure can greatly reduce the propagation of electromagnetic radiation between electronic circuits and the environment. This aids immunity to external interference and reduced electromagnetic emissions. The *shielding effectiveness* of an enclosure depends on the frequency of the electromagnetic radiation, the electromagnetic properties of the material from which it is made, the geometry of the enclosure, and its contents.

Figure 8 shows the shielding effectiveness of a sealed enclosure of relatively low conductivity (carbon-fiber-reinforced plastic). It has a large electric field shielding at all frequencies. Its magnetic field shielding effectiveness is poor at low frequencies; low-frequency magnetic

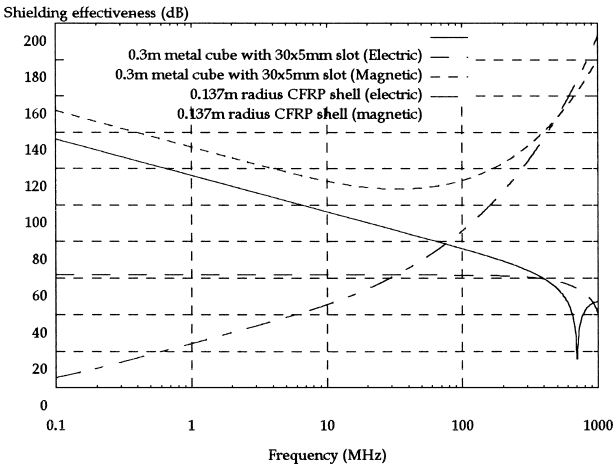


FIGURE 8 Shielding effectiveness at the center of a sealed carbon-fiber reinforced plastic (CFRP) composite enclosure compared with a metal enclosure (cube) of the same volume with an aperture (computed by approximate methods).

shielding is difficult to achieve without a good conductor and/or magnetic materials. In both electric and magnetic cases the shielding increases rapidly at high frequencies due to the skin effect; current from external fields flows only on the outside of the enclosure, causing no disturbance internally (and vice versa for currents due to internal fields). A sealed metal enclosure would have a larger shielding effectiveness. The shielding effectiveness of a metal enclosure with an aperture is also shown in Fig. 8. The aperture dominates the screening of this metal enclosure: electromagnetic energy can more easily pass through the aperture with increasing frequency. The resonant behavior of the metal enclosure can be seen at 700 MHz. This can result in a field enhancement in the enclosure. The analytical solution used for the sealed enclosure does not include the effects of resonances which will be present in real enclosures. In practical metallic enclosures with apertures and unshielded cable penetrations a shielding effectiveness of about 20 dB is typical. It should be stated that the fields within an enclosure can vary considerably with position, so that a value given at a single measurement point is of limited use.

IV. DESIGN FOR ELECTROMAGNETIC COMPATIBILITY

A. The Design Process

Electromagnetic compatibility is affected by almost every aspect of the design and construction of a piece of equipment. It is therefore necessary to integrate EMC considerations into every stage of the design. Figure 9 shows an idealized view of the design process. Design rules for EMC can be applied from the first concept; as the design process continues, rules become hard constraints which may be determined by factors other than just EMC. At each stage in the design some estimate or prediction of EMC must be made. Eventually the system must be tested to see if it meets its EMC specification. Failure to meet the specification must result in redesign until the specification is met.

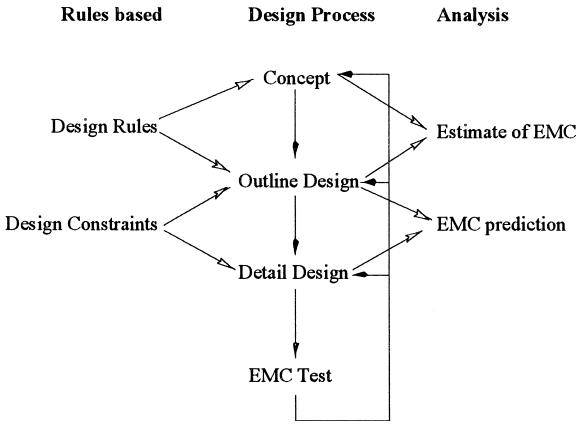


FIGURE 9 An idealized view of the design process.

B. Design Rules and Constraints

Design rules encapsulate a range of measures that are thought to improve the EMC of a system. Often the amount of improvement is difficult to quantify and may vary depending on the details of the system under consideration.

1. Design Concept

If the EMC implications are considered as the design concept is developed, then areas where EMC is an important consideration can be highlighted and an EMC control plan formulated. Alternatives can be considered where EMC weaknesses are suspected. We suggest the following rules.

a. Partition the system into noisy, quiet, robust, and susceptible parts; each can then be considered separately (though the robust can be placed with the noisy, and the quiet with the susceptible).
b. Select internal and external interfaces to minimize emissions and susceptibility (i.e., use the largest signals and narrowest bandwidths possible in circuits that may be susceptible to interference, and use the smallest signals and narrowest bandwidths in circuits that may generate interference).
c. Consider where filtering and shielding are required.
d. Plan the monitoring of EMC throughout the design and development process.

2. Robust, Quiet Circuits

If circuits can be made robust in the presence of interference, then the need for shielding and filtering can be reduced. We suggest the following rules.

a. Select logic circuits with the lowest bandwidth and highest noise margins.
b. Minimize the bandwidth of analog circuits.
c. Apply adequate decoupling on analog and digital circuits.
d. Consider the recovery of analog circuits from transients (simple measures such as limiter diodes can reduce recovery times drastically).
e. Consider carefully partitioning and noise propagation in power supplies.
f. Ensure that unused states in digital (and microprocessor) circuits have transitions into safe

states to allow recovery after disruption by interference and use "watchdog" circuits to force reset after failure in microprocessor systems.
g. Separate I/O busses from the main processor bus to reduce interference transfer to and from interfaces.
h. Use filters and/or isolation to prevent interference propagation.

3. Robust Software

a. Provide system integrity checks (e.g., error detection on code and data).
b. Check peripheral inputs for sensible values (e.g., reject transient changes caused by interference).
c. Check and/or reinitialize peripheral devices periodically to allow recovery from EMI-induced failures.
d. Ensure unused interrupt vectors and unused memory are initialized to cause predictable operation if accessed as a result of interference-induced errors.

4. Quiet Software

- Minimize unnecessary activity (e.g., poll/update interfaces only when necessary, use interrupts to detect changed conditions rather than polling, halt the processor when not active).

5. Physical Layout

a. Partition circuits to minimize propagation of interference from noisy circuits and to susceptible circuits.
b. Provide nearby return for each power and signal connection (by use of power/ground-planes, twisted pairs, shielded cables, etc.).
c. Minimize the physical size of critical circuits to minimize radiation/pickup.
d. Ensure proper termination of cable screens (the screen and enclosure should form a continuous volume in which the conductors are contained; pigtails should not be used).
e. Minimize the dimensions of any apertures and seams in shielded enclosures (the interference propagation depends on the largest dimension; many small holes are better than a single large hole, seams must have good electrical connection avoiding long gaps).

C. Analysis

Analytical techniques for the solution of electromagnetic problems are complex and applicable only to very simple geometries. This has made the direct analytical solution of real EMC problems nearly impossible. However, approximate analysis can often provide useful insight into the magnitude of potential problems and relative performance of possible solutions. With the advent of cheap desktop computing, the evaluation of complex analytical approximations can be achieved in seconds.

D. Computer-Aided Design

Computer-aided design (CAD) tools have permeated much of engineering design and are well established in areas such as circuit analysis and the design of electrical machinery, but are still new to electromagnetic compatibility analysis.

1. Numerical Electromagnetic Solvers

A numerical solution of the electromagnetic properties of arbitrary geometries is possible in principle. In practice the large computational resources required prevent the solution of problems as complex as the prediction of electromagnetic compatibility of complete electronic systems. Numerical methods are widely used to solve simplified geometries in order to allow a better understanding of EMC problems.

2. Signal Integrity

Signal integrity is one area where the use of CAD is well established. Many commercial tools are available for the prediction of signal propagation and crosstalk on printed circuit boards.

3. Design Rules Checking

Automated checking of design rules is an area where CAD can help improve the ease of design. One example is the checking of design rule compliance in printed circuit board layout. A PCB may have many thousands of tracks across six or more layers, making manual checking a slow and error-prone process. Automatic design-rule-checking software is commercially available and can be used to check manufacturing, signal integrity, and EMC rules.

4. Knowledge-Based Systems

Knowledge-based systems attempt to encapsulate the knowledge of a human expert in a computer package. Commercial products which provide design advice and diagnosis of EMC problems are available.

5. Design Frameworks

The difficulty of fully predicting the EMC performance of electronic systems has led to the concept of a design framework which can be used to combine a range of information

on a system which improves in quality as the design and prototype construction progress. At the concept stage a rough estimate of the EMC performance of the system can be obtained by the use of past data, approximate/analytical solutions, and numerical models. This may be enhanced by measurements on subsystems or more-detailed numerical models as the design progresses.

V. ELECTROMAGNETIC COMPATIBILITY REGULATIONS AND STANDARDS

Here we review the EMC regulations and standards of the United States and Europe. Other countries and areas using or adopting EMC regulations include Japan, Australasia, and Taiwan.

A. EMC Regulations

1. Rationale

EMC regulations exist to enforce control of the unwanted emissions from electrical or electronic equipment. This controls 'pollution' of the radio spectrum and provides an environment for the reliable operation of all electrical or electronic equipment.

2. Federal Communications Commission (FCC) Regulations

The FCC administers the use of the radiofrequency spectrum in the United States. Title 47 of the code of Federal Regulations covers telecommunications and contains in five volumes the intentional and incidental use of the spectrum. The parts relevant to EMC are contained in Chapter 1: Part 15 Radio Frequency Devices and Part 18 Industrial Scientific and Medical Equipment.

Part 15 governs emissions from intentional and unintentional radiators and sets out the regulations, technical specifications, and administrative requirements to enable equipment to be marketed without an individual license. Subpart A is concerned with digital devices, subpart B with unintentional radiators, and subpart C with intentional radiators. The FCC classifies equipment into Class A and Class B. Essentially Class A equipment is intended for use in an industrial or commercial environment, while Class B is intended for the residential environment. Accordingly, verification tests for Class A devices are performed by the manufacturer and retained on file; certification by the FCC is not required. For Class B devices FCC certification must be obtained; this is achieved by examining a manufacturer's test results.

The technical requirements for the emission limits are laid down for both conducted emissions and radiated emissions. The methods of measurement are defined by the American National Standards Institute (ANSI) standard C63.4 Methods of Measurement of Radio-Noise Emissions from Low Voltage Electrical and Electronic Equipment in the Range 9 kHz to 40 GHz. The emission limits and the ANSI test methods are derived from CISPR 22 (see Section V.B). Where the device's highest internally generated frequency is greater than 1 GHz, the highest emission frequency to be measured is determined as five times this frequency.

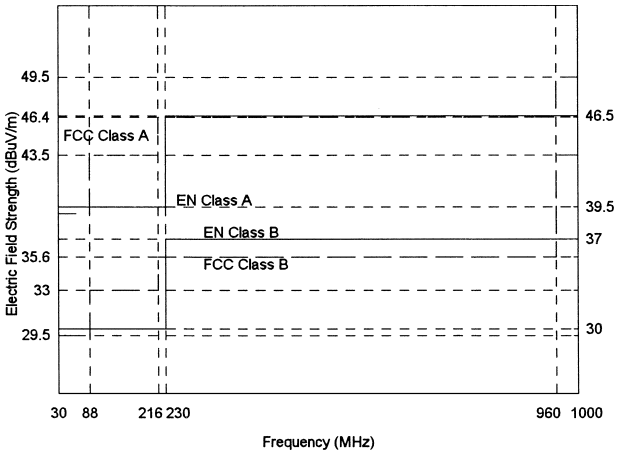


FIGURE 10 The FCC and Euronorm radiated emissions limits measured at 10 m.

Part 18 covers equipment designed to generate and locally use radiofrequency (RF) energy at frequencies greater than 9 kHz for industrial, scientific, and medical (ISM) purposes. It also includes microwave ovens. ISM frequencies are defined at the international level by the International Telecommunications Union (ITU). These frequencies are then allocated at a national level by the national authorities; in the United States the frequencies are allocated by the FCC and are listed in Part 18. Limits and measurements broadly follow CISPR 11 (see Section V.B). Most ISM equipment is subject to FCC certification.

The FCC regulations exclude most industrial electronics equipment.

3. European EMC Regulations—EMC Directive (89/336/EEC)

EMC regulations apply throughout Europe and have had a major impact on the development of EMC regulations throughout the world.

The European regulations result from European Commission Directive 89/336/EEC, which affects all electrical or electronic systems or products sold throughout the European Economic Area (EEA). It also encompasses all electromagnetic phenomena. As a "new approach

directive," the technical requirements are defined by European standards. The new approach directives were designed to remove technical barriers to trade within the European community.

The essential protection requirements of the EMC Directive are as follows:

- Equipment should be constructed so that it will not affect broadcast services or the intended function of other equipment—*the emission aspect*.
- Equipment should have an inherent immunity to externally generated electromagnetic disturbances—*the immunity aspect*

Note that the FCC regulations and Japanese Voluntary Council for the Control of Interference (VCCI) requirements do not have an equivalent requirement for immunity.

The EMC Directive specifies the routes available to manufacturers to show that their product complies with these protection requirements.

The simplest route is to demonstrate compliance with an appropriate European standard. This is a standard whose reference number has been published in the *Official Journal of the European Communities* (OJEC) and is a CENELEC (the European electrical standards body) Euro Norm (EN) that has been transposed into a national standard. An example is EN 55022 (the same as CISPR 22), the emission standard for information technology equipment; the transposed UK standard is BS EN 55022 and the transposed German standard is DIN EN 55022. The standards define emission limits, immunity levels, and the tests that should be performed on equipment to show that it meets these limits and levels. While the European regulations do not explicitly require a product to be tested in order to demonstrate compliance with the protection requirements, it must be demonstrated that the product complies with the standard and therefore by implication must be tested in accordance with the standard. Testing may be performed by the manufacturer or by a third party. There is no requirement for the testing laboratory to have accreditation; however, the use of an accredited laboratory will provide a manufacturer with an assurance that the testing has been performed to the standard correctly and the manufacturer can obtain an accredited test certificate.

When standards are not available to a manufacturer or the equipment has features that mean that a standard can only be partly applied, then the manufacturer must use the "technical construction file" (TCF) route to compliance. Essentially the manufacturer assembles the technical information demonstrating that the product meets the protection requirements. These data, which is likely to include test results, must be reviewed by a competent body appointed by the national authorities. The requirements for a competent body are laid down in Annex II to the EMC Directive. A competent body must demonstrate that it has the appropriate expertise, operates systems that ensure client confidentiality, and has the independence to make an impartial judgement. Such systems are usually ensured by quality assurance to standards such as ISO 9002 and EN45011.

The essential features of a TCF are:

- Part I: description of the apparatus
  a. Identification of the apparatus
  b. A technical description
- Part II: Procedures used to ensure conformity of the apparatus to the protection requirements:
  a. Technical rationale
  b. Detail of significant design aspects
  c. Test data
- Part III: Report (or certificate) from a competent body

For radio transmission equipment (including transceivers) compliance with the Radio and Telecommunications Terminal Equipment (R&TTE) Directive is required except in the case of air traffic management equipment, which is required to conform with the EMC Directive by the "type examination" route. This means that the equipment must be submitted to a notified body (NB; an organization which has been notified to the European Commission by the national competent authority). The NB will require a type examination to be performed. This may be carried out by the NB or one of the NB's approved test laboratories.

When conformance with the protection requirements of the EMC Directive has been demonstrated by one of these three methods, a Declaration of Conformity is issued by the manufacturer and the European Community mark, the CE marking, affixed to the product or its packaging. It should be noted that the CE marking implies that the product complies with all of the new approach directives applicable to it (e.g., machinery safety).

The Australian EMC Framework follows broadly the same pattern as the European regulations, while the U.S. FCC regulations are much more specific and apply to the emission aspects only of "digital" and "industrial scientific and medical equipment," Parts 15 and 18, respectively of the Code of Federal Regulations (CFR) 47 (see Section V.A.1).

B. Overview of Standards

1. Standards Rationale

In order to achieve electromagnetic compatibility between electrical/electronic apparatus, it is necessary to control

(a) emissions from equipment and (b) the level of immunity of equipment to such emissions. This is achieved by using guidelines published as *standards*, which may be enforced by regulations.

Most standards follow the recommendations of the Comité International Special des Perturbations Radioelectriques (or International Special Committee on Radio Interference), CISPR, for establishing emission limits, susceptibility levels, and test procedures. CISPR is a committee of the International Electro-technical Commission (IEC).

Examples of CISPR recommendations are:

- CISPR 11: Limits and methods of measurement of radio disturbance characteristics of industrial, scientific and medical (ISM) radiofrequency equipment (example equivalents: FCC Part 18 and EN 55011).
- CISPR 22: Limits and methods of measurement of radio interference characteristics of information technology equipment (example equivalents: FCC Part 15, ANSI C63.4, EN 55022, and the Japanese VCCI requirements).

2. Relevant Standards for Conformance with EU EMC Regulations

a. General. The EMC Directive defines two methods for demonstrating compliance with the protection requirements. With self-certification, the manufacturer is able to declare that apparatus *conforms to relevant standards*. Alternatively, *a technical construction file* can be prepared, which must include a technical report or a certificate from a competent body.

For manufacturers to *self-certify* their products, they must be designed, built, and tested to meet the requirements of "relevant standards." A "relevant standard" is defined by Article 7 of the EMC Directive as a national standard that has been harmonized with a standard whose reference number has been published in the *Official Journal of the European Communities* (OJEC).

In practice this means that a relevant standard is a Euro Norm (EN), published by CENELEC, the European Committee for Electrotechnical Standardisation.

Euro Norms are derived from CISPR and other IEC publications. It is necessary for individual EEA member states to harmonize their own national standards with the appropriate EN. This means that identical standards will be used in all EEA countries. For example, the British standard that covers emissions from information technology equipment is BS EN 55022. This is harmonized with EN 55 022 and is identical to CISPR 22.

There are two categories of relevant standard: (a) the product, or product family, specific standard, and (b) the generic standard. A product-specific standard applies to a particular type of product or family of products, for example, EN 55 022, which applies to information technology equipment. A product-specific standard takes precedence over generic standards. A generic standard is categorized according to environment type (for example, "residential, commercial, and light industry") and applies to a broad range of product types. Either category of standard may refer to 'reference' or 'basic' standards.

A considerable number of product types have been covered by relevant standards. A representative listing is given in Table I.

b. Generic emission standards. The generic emission standard is EN 50081. Part 1 covers residential, commercial, and light industry environments. Part 2 covers the industrial environment.

Part 1 principally restates the emission limits and test methods defined by EN 55022 Class B, which is the product-specific emission standard for IT equipment; Part 2 does the same with EN 55011, which is the product-specific standard for ISM equipment.

c. Generic immunity standards. The generic immunity standard is EN 50082. Parts 1 and 2 have the same environmental classification as the generic emission standard. These reference the basic standards were introduced by the IEC in the IEC 61000 series and adopted by CENELEC as the EN 61000 series.

Generally, both product-specific and generic standards not only define emission limits and immunity levels, but also specify the test methods to be employed. Manufacturers using these standards to demonstrate compliance with the EMC Directive must be familiar with the contents, and appreciate the implications, of all harmonized EMC standards. In particular they must be aware of sections open to misinterpretation, deficiencies within the standards, and test methods that require significant financial investment.

d. List of representative relevant (harmonised) standards. For manufacturers required to self-certify their products for compliance with the EU EMC Directive, a list of the available product-specific and generic standards is essential. These are made available by the European Commission; an example list is given in Table I.

3. Military Standards

EMC requirements for military equipment have been well understood for many years and as a result the standards

TABLE I Representative List of Harmonized European EMC Standards

Product-specific standards: emission	
EN 50065-1	Mains signaling equipment
EN 55011	Industrial, scientific, and medical (ISM)
EN 55013	Broadcast receivers and associated equipment
EN 55014	Household appliances
EN 55015	Luminaires
EN 55022	Information technology equipment (ITE)
EN 55103-1	Audio, video, audio visual, and entertainment lighting control apparatus for professional use
Product-specific standards: immunity	
EN 55020	Broadcast receivers and associated equipment
EN 55104	Household appliances
EN 50130-4	Alarm systems
Product-specific standards: emission and immunity	
EN 50091-2	Uninterruptable power systems (UPS)
EN 50121	Railway applications
EN 50199	Arc welding equipment
EN 60601-2-3	Electro-medical devices
EN 61131-2	Programmable controllers (PLCs)
Generic standards: emission	
EN 50081-1	Generic class: residential, commercial, and light industry
EN 50081-2	Generic class: industrial
Generic standards: immunity	
EN 50082-1	Generic class: residential, commercial, and light industry
EN 50082-2	Generic class: industrial
Basic standards	
EN61000-3-2	Harmonics
EN61000-3-3	Voltage fluctuation and flicker
EN61000-4-2	ESD
EN61000-4-3	Radiated immunity
EN61000-4-4	EFT/B
EN61000-4-5	Surge
EN61000-4-6	Conducted RF immunity
EN61000-4-8	Power frequency magnetic field immunity
EN61000-4-11	Voltage dips, interruptions

and test methods are well established. DEF Stan 59-41 is the UK MOD standard covering all aspects of EMC from selection of requirements through management of projects to testing and test reporting. The equivalent U.S. standards are MIL-STD-461, which covers EMC requirements, and MIL-STD-462, which covers the test methods.

Effects such as radiation hazards, detection of data from unintentional emissions, and electronic countermeasures are not considered as EMC topics, although related in a number of ways.

The following designations are used in the titles of military standards: R, radiated; C, conducted; MF, magnetostatic field; E, emissions; S, susceptibility (referred to as immunity in commercial standards).

Examples are DEF STAN 59-41, where the first radiated susceptibility test is designated DRS01; and MIL-STD-462, where the first radiated susceptibility test is designated RS101.

Generally the testing methods defined in commercial standards are used in the military standards but the frequency ranges are greater and the severity of susceptibility/immunity test levels is much higher. Examples include MIL-STD-461D RS103, which covers a frequency range of 10 kHz to 40 GHz, compared with EN 61000-4-3, which covers only the frequency range 80 MHz to 1 GHz; and MIL-STD-461D RS103, which requires immunity to a field strength of 200 V/m, compared with EN 61000-4-3, which requires an immunity level of only 10 V/m.

VI. MEASUREMENT AND INSTRUMENTATION

The electromagnetic compatibility of an electronic system can only be fully demonstrated when the system is taken into service. Clearly such a course of action is not acceptable in today's engineering environment and the risks of not achieving EMC must be minimized. Along with the incorporation of EMC in the design process, some attempt must be made to ensure that the equipment will be compatible before it is released into the market. This is achieved by making EMC measurements. Measurements made on a complete system before release to market can be made with reference to EMC standards that quantify the levels of acceptable interference. In addition to these measurements EMC measurements can be made on subsystems bought in from other suppliers or on incomplete parts of the system in order to assess the efficacy of design measures taken before the system is complete. This latter process is particularly important in the design of complex or large systems incorporating electronics such as aircraft or other types of vehicle. Again, relevant standards may be employed.

Many electronic systems can be assessed for EMC in standardized test facilities operated by EMC test houses. There are size limitations on such equipment, but items such as PCs, household appliances, or TV and hi-fi equipment can readily be transported to such facilities. Larger equipment such as motor vehicles are tested in specialist test houses. Very large equipment must be tested after installation, as often the details of the installation can have a bearing on its performance. Normally subsystems of large equipment will have been tested prior to installation.

EMC measurements are inevitably simplified mimics of reality. Consider the hypothetical problem of assessing the interference caused on a computer by a nearby arc welder. The welder is a source of electromagnetic interference energy and the computer is disrupted by that energy. When either item is designed, the interference scenario cannot be predicted in detail. Thus the unwanted interference energy leaving the arc welder must be measured and compared to a predetermined level defined in a standard. This is referred to as an emissions measurement. The interference energy incident upon the computer will cause disruption if the computer is not adequately immune to this interference. Thus, the immunity of the computer to external interference must be measured. The standards are devised such that the required immunity of electronic systems to external interference is greater than the aggregate emissions from neighboring systems. Some systems such as radio transmitters have intentional emissions at energy levels much higher than would ordinarily be allowed. Immunity requirements are adjusted to account for this.

In addition to classifying EMC measurements into emissions and immunity, a further simplifying breakdown is required. The interference between the arc welder and the computer in the above example occurs as a consequence of the transfer of energy between the two. The path that this energy takes may not be immediately apparent. For example, the interference may be a consequence of interference energy conducted away from the welder via the supply mains and entering the computer via its supply. Conversely, it may be due to interference radiated from the welder's leads being picked up by a peripheral lead on the computer. It is not possible to determine each interference scenario in advance, but it can be stated in general that interference energy propagation between the source and the victim is by either conduction or radiation. For this reason, measurements are made for both mechanisms. Consideration of the physics of the energy propagation mechanism leads to a further simplification. In general, conducted interference is a low-frequency phenomenon and radiated interference is a high-frequency phenomenon. This arises because the efficiency of any structure acting as a transmitting or receiving antenna increases with frequency. For significant radiative energy transfer, the "antenna" needs to be comparable to the wavelength in its linear dimensions, i.e., typically more than a tenth of a wavelength. For a system with linear dimensions of 1 m this implies a wavelength of 10 m, corresponding to a frequency of 30 MHz. The boundary is fuzzy. Few emissions measurements are made below 30 MHz and similarly few are made above 100 MHz.

The following sections briefly outline the principal measurement techniques.

A. Emissions

1. Conducted

Conducted emission measurements are made on cables conveying power and signals to and from equipment. Care must be taken to ensure that the signals measured are emerging from the equipment under test (EUT) and are not due to other sources connected to the cable.

Two types of emission are measured. Common-mode emissions measurements use a calibrated current transformer to measure the total interference current present on a cable. The output of the current transformer is fed to a measurement receiver that indicates the voltage present at its input port. The current transformer is calibrated in terms of its transfer impedance. This impedance relates the receiver input voltage to the current flowing on the cable. The EUT can be isolated from the far end of the cable by placing an absorbing ferrite clamp around the cable.

Such a clamp provides a stable absorbing load for the interference currents on the cable and the required isolation from other devices connected to the cable. Often the current transformer is combined in the same structure as the absorbing clamp.

Differential mode emissions measurements are made as voltage measurements between individual pairs of conductors in a cable. Again isolation is required. The isolation is provided by a line impedance isolation network (LISN) inserted into the conductor pair. The simplest form of LISN provides a low-pass filter for the intentional signals on the conductor pair and a barrier for higher frequency interference signals propagating in either direction along the cable. Along with the receiver's input impedance, it provides a stable and defined measurement impedance for the interference signals.

2. Radiated

Radiated emission measurements are made using a defined environment into which the EUT radiates. The current international standard environment is the open-area test site (OATS). The radiation from the EUT is measured using a calibrated antenna and a measurement receiver. The EUT and antenna are positioned on the OATS at the foci of an ellipse. The area of the OATS is defined as the area of the ellipse with a major diameter of twice the focal length and a minor diameter of the square root of three times the focal length. Typical EUT-antenna spacings are 10 and 3 m. A 10-m OATS thus has an elliptical area of 20 by 17.3 m. The signal received by the antenna is the combination of the direct wave from the EUT and the ground reflection. In order to preserve the repeatability of measurements on a given site and the reproducibility of measurements between sites, the ground reflection has to be stabilized against changes introduced by climatic effects. A metallic ground plane is used beneath the antenna and the EUT and in the space between them for this purpose. In order to measure the maximum signal arising from the combination of the two waves, the antenna height is scanned from 1 to 4 m. The OATS suffers from the presence of ambient signals that can mask the emissions from the EUT. This disadvantage can be overcome by enclosing the OATS in a screened room with radio-absorbing material on the walls. Such a facility is called a semi-anechoic chamber. Recently, it has been suggested that a fully anechoic chamber with radio absorber on its floor would be a better environment for measuring radiated emissions. Such a chamber would not need to have the antenna height scan and would be compatible with chambers used for radiated immunity measurements as described below. It remains to be seen if this suggestion will be adopted.

B. Immunity

1. Conducted

Conducted immunity is measured using transducers similar to those used for conducted emissions. Energy is injected onto cables in both common mode and differential mode. Conducted emission measurements are made using receivers tuned across a specified frequency range, the frequency domain. Conducted immunity measurements can also be made in the frequency domain by injecting energy at specified frequencies. Other time-domain waveforms can also be used such as pulses or bursts of pulses in order to simulate known threats to the EUT.

2. Radiated

The radiated immunity of an EUT is measured by illuminating the EUT with a radio wave that simulates the perceived threat. This is always done in an anechoic chamber in order to prevent the radiated energy from causing interference to other systems. In general, the threat to an EUT is likely to come from an intentional radio transmitter, usually a low-power mobile transmitter. The EUT is illuminated at an appropriate field strength by an amplitude-modulated signal the modulation of which mimics the modulation of the threat. For example, in analog amplitude modulation schemes the chosen standard is 80% modulation depth with a 1-kHz tone. GSM mobile phone modulation is simulated by a 217-Hz pulse modulation. Frequency and phase modulation is simulated by a constant-strength carrier. The EUT needs to be observed when under stress. For this reason the threat is applied at a series of frequencies with each having a defined dwell time. The frequency is normally incremented in 1% or 2% steps.

3. Electrostatic Discharge

Electrostatic discharge (ESD) is a further electromagnetic phenomenon that may cause equipment malfunction. The most common scenario is the discharge of a charged human body through a finger onto the EUT. The source of the charge is the triboelectric effect acting on floor coverings and synthetic clothing. Charging potentials of up to 16 kV can be experienced. An electrostatic discharge gun simulates this threat by approximating the charged human body with a series resistor/capacitor circuit with the capacitor charged to an appropriate potential. Typical circuit values for an adult human are 200 pF in series with 200 ohms. The discharge is through an artificial finger either with an air discharge to the EUT or a direct contact discharge. The ESD event results in significant reactive fields in the vicinity of the discharge. A further test requires a discharge to an earthed plate close to the

EUT. The potential disturbance via the reactive field is assessed.

SEE ALSO THE FOLLOWING ARTICLES

ELECTROMAGNETICS • MICROWAVE COMMUNICATIONS • RADAR • RADIO PROPAGATION • RADIO SPECTRUM UTILIZATION • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • WIRELESS COMMUNICATION

BIBLIOGRAPHY

Archambeault, B., Ramahi, O., and Brench, C. (1998). "EMI/EMC Computational Modeling Handbook," Kluwer Academic Publishers, Norwell, Massachusetts.

Department of Trade and Industry (UK). (1992). "The Electromagnetic Compatibility Regulations," HMSO, London.

Department of Trade and Industry (UK). (1992). "Guidance Document on the Preparation of a Technical Construction File As Required by EC Directive 89/336," HMSO, London.

Goedbloed, J. (1992). "Electromagnetic Compatibility," Prentice Hall, Englewood Cliff, NJ.

Hoeft, L. O., and Hofstra, J. S. (1988). "Measured electromagnetic shielding performance of commonly used cables and connectors," *IEEE Trans. EMC* **30**(3), 260–275.

Hubing, T. (1991). "A survey of numerical electromagnetic techniques," *In* "ITEM Update," pp. 17–13, 60, 62, Robar Industries, West Conshohocken, Pennsylvania. URL:www.rbitem.com.

Marshman, C. (1995). "The Guide to the EMC Directive 89/336/EEC," 2nd ed., EPA Press, Saffron Walden, UK.

Molinkski, T. S., Feero, W. E., and Damsky, B. L. (2000). "Shielding grids from solar storms," *IEEE Spectrum* **37**(11), 55–60.

Paul, C. R. (1992). "Introduction to Electromagnetic Compatibility," Wiley Interscience, New York.

Tesche, F. M., Ianoz, M. V., and Karlsson, T. (1997). "EMC Analysis Methods and Computational Models," Wiley, New York.

Williams, T. (1996). "EMC for Product Designers," Newnes, Butterworth-Heineman, Woburn, MA.

Williams, T., and Armstrong, K. (2000). "EMC for Systems and Installations," Newnes, Butterworth-Heineman, Woburn, MA.



Electromagnetics

Sheila Prasad
Northeastern University

I. Historical Introduction
II. Maxwell's Equations
III. Electromagnetic Waves
IV. Applications of Electromagnetics
V. Recent Developments

GLOSSARY

Antenna Structure that is designed is such a way that it will radiate electromagnetic power efficiently.
Charge Fundamental physical quantity that is indestructible and is characterized by mutual interactions with other charges.
Current Time rate of change of charges that are in motion.
Electric field Force per unit charge.
Electromagnetic energy Energy stored in the electromagnetic field.
Elementary dipole Positive and negative charge that are tightly bound together.
Elementary magnet Electron rotating about an axis.
Magnetic field Force produced by an electric current.
Magnetization Orientation of elementary magnets along parallel axes due to an external force.
Phase velocity Speed with which a wave front moves in space.
Plane wave Wave for which constant-phase surfaces are planes perpendicular to the direction of propagation.
Potential Potential energy of the electromagnetic field.
Spherical wave Wave for which constant-phase surfaces are spheres perpendicular to the direction of propagation.
Wavelength Distance between two constant-phase surfaces with a phase difference of 180°.

ELECTROMAGNETICS is the description of the electricity and magnetism that exist in space and various materials. These physical phenomena are described in terms of electric and magnetic fields created from electric charges and currents and forces associated with them. A precise mathematical formulation of these physical effects is given in Maxwell's equations. The energy in the electromagnetic field is transported by electromagnetic waves, which travel in unrestricted space, infinitely large material media, or in physical structures that guide the wave in specific directions.

I. HISTORICAL INTRODUCTION

The history of the development of electromagnetics is the history of the development of electrical science. Abstract mathematical theory was applied to the description

of physical phenomena, and this eventually evolved into modern technology, which is continually changing.

The history of electromagnetism may be treated in terms of the three periods of its growth. In the first, the fundamental concepts of action at a distance between charges and currents were developed. Earlier, it was proposed that action could take place only by contact through a material medium. This concept is as old as the history of man. Aristotle was a believer in such action, and this same idea was propounded by the philosophers of the East who predated Aristotle. Much later, even Newton considered the idea that one body could act on another through empty space an impossible one. It was much easier for these philosophers to explain the process of throwing a stone, which implied action by contact, than the mere falling of a stone due to the interaction of it and the earth with no visible push.

It was Newton, however, who made the concept of action at a distance acceptable. His law of gravitation gave the force between two masses at a distance without referring to a mechanical medium. The mathematician Euler attempted to explain the theories of gravitation, light transmission, and the interaction of permanent magnets in terms of the intervening material medium, the ether. Newton's inverse square law of gravitation formed the basis of early work on the interaction between charges at a distance in dependent of any intermediate material in contact. The inverse square law for electric charges was first suggested by Priestly (1766), who used an electrometer, and was discussed by Cavendish (1771), and the formulation familiar to us today was given by Coulomb (1785), who carried out experiments using a torsion balance. The inverse square law for magnetic poles was expressed by Michell (1750) for the first time. Much later (1820–1821), the magnetic effects of currents were investigated by Oersted, Biot, Savart, and Faraday. During this time, Laplace formulated a law of action at a distance between elements of current and magnetic dipoles. Ampere (1823) performed experiments that led to the law of force between current elements, which was, once again, a law of action at a distance. Ohm's law (1826) was followed by Faraday's law of induction (1832) and Lenz's law (1834). Potential theory was expounded by Gauss and Green separately during this same period. Neumann and Weber (1845–1847) expounded their work on induction resulting from current-carrying conductors in motion and due to the rise and decay of currents. The current and voltage laws that are the basis for electrical engineering were given by Kirchhoff (1845). The work of developing a fundamental law of electromagnetic action at a distance was continued by Grassmann, Riemann, and Clausius during this period. The idea of the propagation of this action was first suggested by Gauss. This was extended by Riemann (1858), who showed that such propagation moved with a velocity equal to that of the velocity of light. The last important work of this period was by Lorenz (1867), who suggested using retarded scalar and vector potentials. This showed clearly that it was not necessary to have contact with a medium to have action at a distance.

The second period of growth was marked by the outstanding work of Maxwell, which laid the foundation of electromagnetics. Kelvin (1847) attempted to explain the results of his electrical experiments with theories of elasticity. The ideas of Kelvin and Faraday on electromagnetic force were the basis for Maxwell's investigations (1864). However, Maxwell's entire hypothesis was based on an all-pervading mechanical medium, the ether. The field equations formulated by Maxwell govern all electromagnetic phenomena and form the basis of an understanding of electromagnetics. The equations were so comprehensive that they included all the earlier observations: the laws of Coulomb, Gauss, Faraday, and Ampere. Maxwell's formulation was simplified by Heaviside and Hertz. The flow of energy in the ether was proposed by Poynting (1884) to be governed by a vector that now bears his name. Hertz (1887) demonstrated the existence of electromagnetic waves in the ether. Radio transmission was achieved for the first time, and Maxwell's theories were verified.

The third and final period of the growth of electromagnetic theory involved the development of the theory of retarded action at a distance between charges and currents. Lorentz (1895) coordinated the earlier theories of action between charges and currents with Maxwell's general theory of the state of the ether. He theorized that matter contains electrons that act on each other in various ways to produce all electromagnetic (including optical) effects. Lorentz assumed that the electromagnetic field characterizes and is propagated by the ether.

The ether was proposed as a means of transporting electrical effects from one charge to another rather than the basis of all electromagnetic phenomena as proposed by Maxwell. The theory of relativity laid to rest all claims about the legitimacy of the ether hypothesis, and the conclusion was that there is no ether. Maxwell's theory of the electromagnetic ether is only of historical significance. However, the field equations of Maxwell continue to be the basis of macroscopic electromagnetic theory. The fundamental law of macroscopic electromagnetism as expressed in the field and force equations is interpreted as retarded action at a distance.

II. MAXWELL'S EQUATIONS

A. The Density Functions

Electric charge is the fundamental physical quantity from which all other concepts in electromagnetics are derived.

It is indestructible in that it can be neither created nor destroyed—this is the principle of conservation of charge. The electrodynamical model, which serves as the mathematical foundation for electromagnetics, depends on continuous functions, which are called densities or density functions. These depend on and take account of the magnitude, the distribution, and the relative velocities of the charges. These density functions are termed the volume density of charge ρ and the surface density of charge η when the statistically stationary state is being considered. In this state, the charges are all in random motion. The statistical behavior of the charges in motion is obtained by taking a time average; from the statistical point of view, a volume containing millions of charges moving randomly is indistinguishable from the same volume containing the same charges with each fixed at an average rest position. Hence, the statistically stationary state is essentially a static state. The volume density of charge ρ describes the average condition of total charge throughout the region under consideration but does not describe the average separation and orientation of the statistical rest positions of positive and negative charges. When a region consisting of tightly bound charges (positive and negative) is exposed to an external force that attracts negative and repels positive charges, the negative charge is pulled away from the positive charge, and the separation of the two charges is oriented in the same direction as the external force. Such a structure is called a dipole. If the positive charge q is separated a distance d from an equal negative charge, then the average polarization or dipole moment is given by

$$\mathbf{p} = q\mathbf{d}, \tag{1}$$

where $\mathbf{d}$ is a vector drawn from the center of the negative charge to the center of the positive charge. The volume density of polarization (also called the polarization) is denoted by $\mathbf{P}$. It is the polarization per unit volume. Since the volume function $\mathbf{P}$ is a measure of the average density of polarization vectors due to individual dipoles in a small region about a Point, the component of $\mathbf{P}$ directed along the outward or external normal to a closed surface at any point,

$$P_{\mathrm{n}} = \hat{\mathbf{n}} \cdot \mathbf{P} \tag{2}$$

($\hat{\mathbf{n}}$ is the unit external normal), gives the average sum of the outwardly directed normal components of the polarization vectors piercing the unit area of the surface on which P_{n} is defined. Then

$$\int_{\Sigma} \hat{\mathbf{n}} \cdot \mathbf{P}\, d\sigma \tag{3}$$

(where Σ is a surface and $d\sigma$ the element of surface) measures the number of polarization vectors piercing the surface normally as well as the total positive charge leaving (or negative charge entering) the volume enclosed by the surface. Hence, the net addition of positive charge per unit volume (or removal of negative charge per unit volume) is given by

$$-\int_{\Sigma} \hat{\mathbf{n}} \cdot \mathbf{P}\, d\sigma/\Delta\tau, \tag{4}$$

Where $\Delta\tau$ is the volume enclosed by Σ. Then

$$\lim_{\Delta\tau\to 0} -\int_{\Sigma} \hat{\mathbf{n}} \cdot \mathbf{P}\, d\sigma/\Delta\tau = -\boldsymbol{\nabla} \cdot \mathbf{P}\,(\text{divergence of } \mathbf{P}). \tag{5}$$

Hence the diveragence represents the total outward normal flux of the polarization and measures the charge added.

The steady state is a generalization of the static state. It is characterized by a steady drift or circulation of electric charges relative to the statistically stationary rest positions that characterize the static condition. A steady average flow is assumed to be superimposed on the random motions of the charges. A steady drift might consist of one kind of charge flowing in a definite direction or of two kinds of charges flowing in opposite directions. Such a drift is called a convection current. A special form of the convection current is the steady drift of electrons relative to statistically stationary nuclei. This is called a conduction current. The volume density of the moving charge or the volume density of the convection or conduction current denoted by $\mathbf{J}$ is a measure of the average drift of electric charges both in magnitude and direction. If there is a layer of free electrons moving with a steady drift velocity on the surface, a surface density of convection current may be defined.

In the steady state, the model consisting of electrons rotating about an axis through an atom is an elementary magnet. It is produced by forces causing the electrons in an atom to change their random orbits and circulate about a common axis. Such elementary magnets are oriented along parallel axes with a common direction of rotation, and this orientation is called the magnetization due to circulation, $\mathbf{m}_{\mathrm{c}}$. The magnetization per unit volume or the volume density of magnetization of the circulating electrons is denoted by $\mathbf{M}_{\mathrm{c}}$. There is also a magnetization due to the spin of the electrons. The spin is the property of the electron whereby it has an intrinsic angular momentum in addition to the angular momentum of its orbital motion. The spin magnetization is denoted by $\mathbf{m}_{\mathrm{s}}$ and the volume density of spin magnetization is denoted by $\mathbf{M}_{\mathrm{s}}$. The volume density of spin magnetization $\mathbf{M} = \mathbf{M}_{\mathrm{c}} + \mathbf{M}_{\mathrm{s}}$.

The magnetization vector $\mathbf{M}$ is a continuous function that measures the average density and direction of magnetization vectors in a small region about any point. Since it is parallel to the axis of rotation of the circulating charges,

$\mathbf{M}$ is perpendicular to the plane of motion of the charges to be represented by it. The vector $\hat{\mathbf{n}} \times (-\mathbf{M})$ is perpendicular to $\mathbf{M}$, and $\hat{\mathbf{n}}$ and is proportional to $\mathbf{M}$ (the negative sign preserves the right-hand screw convention for the rotating current whirls of positive charge). The magnitude of $\hat{\mathbf{n}} \times (-\mathbf{M})$ is a measure of the average vector sum of the tangential components of the magnetization vectors.

The expression

$$\int_{\Sigma} \hat{\mathbf{n}} \times \mathbf{M}\, d\sigma \tag{6}$$

is a measure of the average number of positive half-current whirls that are cut and added to the volume enclosed by Σ. Hence the current per unit volume appearing in $\boldsymbol{\Delta}\tau$ is

$$\int_{\Sigma} \hat{\mathbf{n}} \times \mathbf{M}\, d\sigma / \Delta\tau. \tag{7}$$

So, it represents the mean density of magnetization current in $\Delta\tau$. Then

$$\lim_{\Delta\tau \to 0} \int_{\Sigma} \hat{\mathbf{n}} \times \mathbf{M}\, d\sigma / \Delta\tau = \boldsymbol{\nabla} \times \mathbf{M}\,(\text{curl of } \mathbf{M}). \tag{8}$$

The actual effective volume and surface densities of charge are given by

$$\begin{aligned} \bar{\rho}(r) &= \rho(r) - \boldsymbol{\nabla} \cdot \mathbf{P}(r), \\ \bar{\eta}(r) &= \eta(r) + \hat{\mathbf{n}} \cdot \mathbf{P}(r). \end{aligned} \tag{9}$$

The effective volume and surface densities of current are

$$\begin{aligned} \bar{\mathbf{J}}(r) &= \mathbf{J}(r) + \boldsymbol{\nabla} \times \mathbf{M}(r), \\ \bar{\mathbf{K}}(r) &= \mathbf{K}(r) - \hat{\mathbf{n}} \times \mathbf{M}(r). \end{aligned} \tag{10}$$

The static and steady states are both stationary since the density functions are independent of time. In the nonstationary state, the same density functions may be used to describe the instantaneous distributions but vary in time at every point. Hence,

$$\partial \bar{\rho} / \partial t = \partial \rho / \partial t - \boldsymbol{\nabla} \cdot \partial \mathbf{P} / \partial t. \tag{11}$$

If $\Delta\tau$ is a volume cell enclosed by a surface Σ, the total charge is $\rho \Delta\tau$, and the rate of increase of the positive charge is given by the time derivative of it. This must be equal to the net positive charge entering across $\Delta\tau$, per unit time due to energy conservation. The resulting equation is

$$\partial \rho / \partial t = \lim_{\Delta\tau \to 0} -\int_{\Sigma} \hat{\mathbf{n}} \cdot \mathbf{J}\, d\sigma / \Delta\tau = -\boldsymbol{\nabla} \cdot \mathbf{J}, \tag{12}$$

which is the equation of continuity for electric charge. With Eq. (11) and the vector identity $\boldsymbol{\nabla} \cdot \boldsymbol{\nabla} \times \mathbf{M} = 0$, the more general equation of continuity is obtained as

$$\partial \bar{\rho} / \partial t + \boldsymbol{\nabla} \cdot \overline{\rho_{\mathrm{m}} \mathbf{v}} = 0, \tag{13}$$

where the essential volume density of moving charge is defined as

$$\overline{\rho_{\mathrm{m}} \mathbf{v}} = \mathbf{J} + \boldsymbol{\nabla} \times \mathbf{M} + \partial \mathbf{P} / \partial t. \tag{14}$$

The general surface equation of continuity is

$$\partial \bar{\eta} / \partial t + \boldsymbol{\nabla} \cdot \overline{\eta_{\mathrm{m}} \mathbf{v}} - \hat{\mathbf{n}} \cdot \overline{\rho_{\mathrm{m}} \mathbf{v}} = 0, \tag{15}$$

where $\bar{\eta}$ and $\overline{\rho_{\mathrm{m}} \mathbf{v}}$ are defined in Eqs. (9) and (14) and $\bar{\eta}_{\mathrm{m}} \bar{\mathbf{v}}$ is the surface density of moving charge or current defined in Eq. (10), and

$$\overline{\eta_{\mathrm{m}} \mathbf{v}} \equiv \bar{\mathbf{K}} = \mathbf{K} - \hat{\mathbf{n}} \times \mathbf{M}.$$

B. Maxwell's Equations

The density fields of matter described in Section II.A have to be interconnected, and that is the fundamental purpose of the mathematical description of space in terms of its electromagnetic properties. In such a description, space consists of a coordinate system that assigns three coordinates to every point to provide a relationship to an arbitrarily selected region. In certain regions, which are located by these coordinates, the density fields characterizing matter have nonzero values. These regions define the positions of the mathematical bodies in terms of the coordinates. In empty space the density fields are zero. The mathematical model includes all of space in order to interconnect the different density fields that are scattered over it. This is done by assigning two vectors to every point in space whether it is empty or has nonzero density fields. The electrical structure of mathematical space is described in terms of two vector fields: the electric vector $\mathbf{E}$ and the magnetic vector $\mathbf{B}$. An electric field is said to exist in a region in which $\mathbf{E}$ has a value at every point. A magnetic field exists in a region where $\mathbf{B}$ has a value at every point. The superposition of the two fields is called the electromagnetic field. Thus, the mathematical description of the structure of space is completely identified with the electromagnetic field. The definition of each of the two vectors $\mathbf{E}$ and $\mathbf{B}$ involves a numerical, experimentally determined proportionality constant with appropriate dimensions. These are the fundamental electric constant ε_0 called the permittivity of free space and the fundamental magnetic constant μ_0 called the permeability of free space. These factors are necessary to get the numerical coordination between the mathematical model of electromagnetism and experimental measurements.

The definition of the vectors $\mathbf{E}$ and $\mathbf{B}$ in terms of the density fields that characterize the space occupied by matter depends on a fundamental theorem of vector analysis. This theorem states that a vector field is uniquely determined if its divergence and curl are specified and if the normal component of the field is known over a closed

surface or if the vector vanishes as $1/r^2$ at infinity, where r is the distance from the density distribution.

The definition of the vectors **E** and **B** in terms of their respective divergences and curls is the second fundamental principle of electromagnetics. The first fundamental principle is the conservation of charge, which was expressed mathematicaly in the equation of continuity. The second principle (which contains the first) is expressed by a set of partial differential equations, called Maxwell's equations. These express the divergence and curl of the **E** and **B** vectors in terms of the density functions and the constants ε_0 and μ_0 as follows:

$$\varepsilon_0 \nabla \cdot \mathbf{E} = \bar{\rho}, \tag{16}$$

$$\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t, \tag{17}$$

$$\mu_0^{-1} \nabla \times \mathbf{B} = \overline{\rho_m \mathbf{v}} + \varepsilon_0 \, \partial \mathbf{E}/\partial t, \tag{18}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{19}$$

It is assumed that the region (or regions) that is characterized by $\bar{\rho}$ is as a whole at rest relative to the observer. The defining relations, Eqs. (16)–(19), completely describe the electromagnetic field in terms of the essential volume characteristics. The vectors **E** and **B** define a macroscopic field.

In the stationary states, all of the functions are constant in time, and the field equations are of the form

$$\varepsilon_0 \nabla \cdot \mathbf{E} = \bar{\rho}, \tag{20}$$

$$\nabla \times \mathbf{E} = 0, \tag{21}$$

$$\mu_0^{-1} \nabla \times \mathbf{B} = \bar{\mathbf{J}} = \mathbf{J} + \nabla \times \mathbf{M}, \tag{22}$$

$$\nabla \cdot \mathbf{E} = 0. \tag{23}$$

The unit for the electric vector **E** is volts per meter, and the unit for the magnetic vector **B** is webers per square meter or teslas. The numerical values of the universal constants ε_0 and μ_0 are obtained from standard experiments to be

$$\varepsilon_0 = 8.854 \times 10^{-12} = 1/36\pi \times 10^{-9} \text{ F/m}$$

and

$$\mu_0 = 1.257 \times 10^{-6} = 4\pi \times 10^{-7} \text{ H/m},$$

respectively.

C. Field Equations at a Surface: Boundary Conditions

A boundary surface is defined to be either the mathematical envelope between a charged region and empty space, where the density fields associated with the region vanish, or the mathematical envelope between two electrically different regions in contact, where the density fields associated with the two change abruptly. Conditions at the boundary between a charged region and space are obtained from those for two charged regions in contact by setting one set of density fields equal to zero.

Since the electromagnetic vectors **E** and **B** are defined in terms of all of the volume densities, they cannot represent more rapid fluctuations in electrical conditions than can the densities themselves. Therefore, discontinuities in **E** and **B** can exist only at a boundary where an abrupt change from one set of densities to another occurs.

Maxwell's equations, written for surface effects on the boundary between two regions, have the form

$$\varepsilon_0 \hat{\mathbf{n}}_1 \cdot \mathbf{E}_1 + \varepsilon_0 \hat{\mathbf{n}}_2 \cdot \mathbf{E}_2 = -(\eta_1 + \eta_2 + \hat{\mathbf{n}}_1 \cdot \mathbf{P}_1 + \hat{n}_2 \cdot \mathbf{P}_2) \tag{24}$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}_1 + \hat{\mathbf{n}}_2 \times \mathbf{E}_2 = 0, \tag{25}$$

$$\mu_0^{-1}(\hat{\mathbf{n}}_1 \times \mathbf{B}_1 + \hat{\mathbf{n}}_2 \times \mathbf{B}_2) = -(\mathbf{K}_1 + \mathbf{K}_2 - \hat{\mathbf{n}}_1 \times \mathbf{M}_1 - \hat{\mathbf{n}}_2 \times \mathbf{M}_2) \tag{26}$$

$$\hat{\mathbf{n}}_1 \cdot \mathbf{B}_1 + \hat{\mathbf{n}}_2 \cdot \mathbf{B}_2 = 0, \tag{27}$$

where $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ are the exterior unit normals (pointing *out* of the region in each case). The boundary conditions can be interpreted easily.

The relations (24) and (27) apply to the normal components of the vectors **E** and **B**. Thus, Eq. (24) states that the normal component of the electric vector is discontinuous in crossing a boundary surface, and the magnitude of the discontinuity is the essential surface characteristic of charge $\bar{\eta}$ divided by ε_0. Equation (27) states that the normal component of the magnetic vector is continuous across all boundaries. The interpretation of Eqs. (25) and (26) is somewhat more involved since the vector product of the external normal to the surface of a region and one of the field vectors actually does not specify any particular component of the field vector. It defines an axial vector that has the *magnitude of the tangential component* of the field vector at the surface and a direction normal to the plane formed by the field vector and the external normal. It follows that the magnitude of the discontinuity of the axial vector so defined in the tangential component of the field vector. Therefore, Eq. (25) requires that the tangential component of the electric vector be continuous in crossing all boundaries, wherease Eq. (26) requires the tangential component of the magnetic vector to be discontinuous by a magnitude equal to the essential surface density of current $\overline{\eta_m \mathbf{v}}$ divided by μ_0. These boundary conditions are illustrated in Fig. 1.

The field equations may also be written in terms of the auxiliary field vectors **D** and **H**, which are defined as

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} \tag{28}$$

and

$$\mathbf{H} = \mu_0^{-1} \mathbf{B} - \mathbf{M}. \tag{29}$$

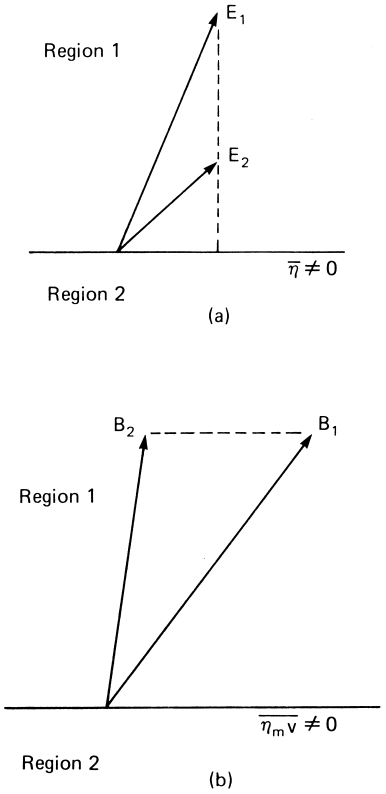


FIGURE 1 Electric (a) and magnetic (b) fields at a boundary.

At all points in space or in bodies where **P** and **M** vanish, $\mathbf{D} = \varepsilon_0 \mathbf{E}$ and $\mathbf{H} = \mu_0^{-1}\mathbf{B}$. Since ε_0 and μ_0 are scalars, the vectors **D** and **H** point in the same direction as **E** and **B** at all points where **P** and –**M** vanish or are not defined.

Maxwell's equations expressed in terms of the auxiliary vectors **D** and **H** are

$$\nabla \cdot \mathbf{D} = \rho, \tag{30}$$

$$\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t, \tag{31}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \partial \mathbf{D}/\partial t, \tag{32}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{33}$$

The corresponding surface equations are

$$\hat{\mathbf{n}}_1 \cdot \mathbf{D}_1 + \hat{\mathbf{n}}_2 \cdot \mathbf{D}_2 = -(\eta_1 + \eta_2), \tag{34}$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}_1 + \hat{\mathbf{n}}_2 \times \mathbf{E}_2 = 0, \tag{35}$$

$$\hat{\mathbf{n}}_1 \times \mathbf{H}_1 + \hat{\mathbf{n}}_2 \times \mathbf{H}_2 = -(\mathbf{K}_1 + \mathbf{K}_2), \tag{36}$$

$$\hat{\mathbf{n}}_1 \cdot \mathbf{B}_1 + \hat{\mathbf{n}}_2 \cdot \mathbf{B}_2 = 0. \tag{37}$$

The boundary conditions are illustrated in Figs. 2 and 3, with $\eta = 0$ and $\mathbf{K} = 0$. Media with **P** parallel to and in the same direction as **E** are said to be *dielectric*. Media with $-\mathbf{M}$ parallel to and in the same direction as **B** are *diamagnetic*. Media with $-\mathbf{M}$ parallel and directed opposited to **B** are *paramagnetic* when **M** is small and *ferromagnetic* when **M** is large.

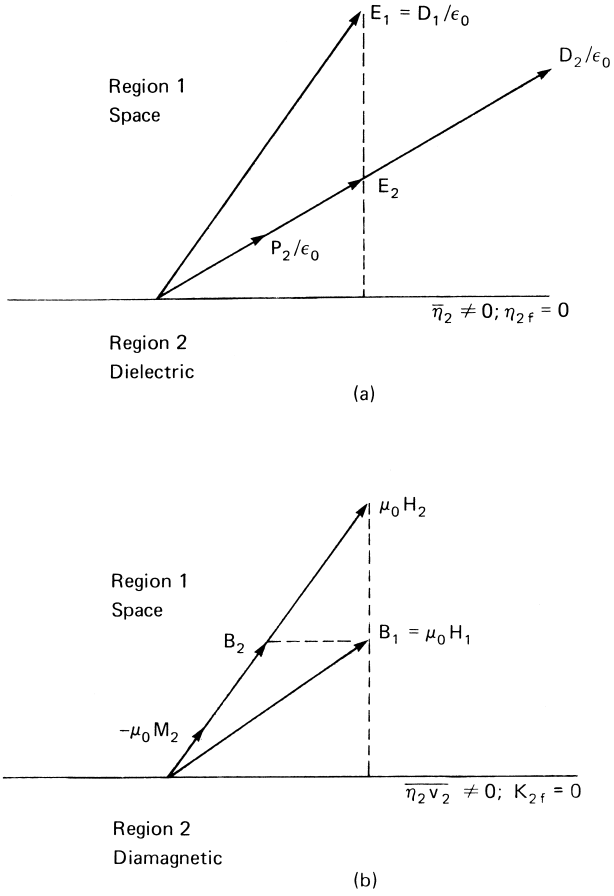


FIGURE 2 Electric (a) and magnetic (b) field vectors at a boundary.

Two universal constants may be derived from ε_0 and μ_0. The characteristic velocity of space (the velocity of light) $c = 1/\sqrt{(\varepsilon_0 \mu_0)} = 3 \times 10^8$ m/sec, and the characteristic resistance of space $\zeta_0 = \sqrt{(\mu_0/\varepsilon_0)} = 120\pi$ ohms.

D. Integral Forms of the Field Equations

Maxwell's equations defining the electromagnetic field consist of four simultaneous partial differential equations. It is possible to transform them into integral relations that are often more convenient in the solution of problems,

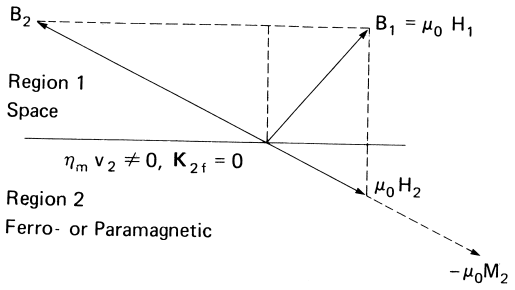


FIGURE 3 Magnetic field vectors at a boundary.

particularly those characterized by symmetry. This is accomplished by using general integral theorems of calculus. The two theorems are the divergence theorem and Stokes's theorem.

The divergence theorem transforms a volume integral into an integral evaluated over a surface enclosing the volume. Let **A** be a continuous vector point function that defines a vector field. If any volume V is chosen in this field, it will be contained in a closed surface S. Let dV be an element of the volume, dS an element of the enclosing surface, and $\hat{n}$ an external normal to the surface. The theorem is

$$\int_V \nabla \cdot \mathbf{A}\, dV = \int_S \hat{\mathbf{n}} \cdot \mathbf{A}\, ds. \tag{38}$$

Stokes's theorem is a theorem for transforming a surface integral over a cap-or cup-shaped surface (a surface that does not enclose a volume) into a line integral around the closed boundary of the surface. Consider any open cap or cup-shaped surface S, which may have any form whatsoever, from a flat disc enclosed by the boundary line s to a deep balloon with only a narrow opening enclosed by the boundary line s. Let this surface be entirely in the field of a continuous vector point function **A**. The theorem is

$$\int_{S(\text{cap})} \hat{\mathbf{n}} \cdot (\nabla \times \mathbf{A})\, dS = \int_{s(\text{closed line})} \mathbf{A} \cdot \mathbf{ds}. \tag{39}$$

The line integration around the closed boundary s is to be performed such that the right-hand screw convention is satisfied with respect to the normal $\hat{n}$ to the surface S.

Stokes's theorem and the divergence theorem may be applied to the four field equations. In integral form they become:

1. Gauss's law:

$$\varepsilon_0 \int_{S(\text{closed})} \hat{\mathbf{n}} \cdot \mathbf{E}\, dS = \bar{Q}. \tag{40}$$

2. Faraday's law:

$$\int_s \mathbf{E} \cdot d\mathbf{s} = -\frac{\partial}{\partial t} \int_{S(\text{cap})} \hat{\mathbf{N}} \cdot \mathbf{B}\, dS. \tag{41}$$

3. The Ampere–Maxwell theorem:

$$\mu_0^{-1} \int_s \mathbf{B} \cdot \mathbf{ds} = \bar{I} + \frac{\partial}{\partial t} \varepsilon_0 \int_{S(\text{cap})} \hat{\mathbf{N}} \cdot \mathbf{E}\, dS. \tag{42}$$

4. Gauss's law:

$$\int_{S(\text{closed})} \hat{\mathbf{n}} \cdot \mathbf{B}\, dS = 0. \tag{43}$$

Here $\hat{\mathbf{n}}$ is the external normal to a closed surface S enclosing the volume V, $\hat{\mathbf{N}}$ is the normal to a cap surface S (open surface) bounded by a closed contour **s**,

$$\bar{Q} = \int_\tau \bar{\rho}\, dV + \int_\Sigma \bar{\eta}\, dS, \tag{44}$$

and

$$\bar{I} = \int_{S(\text{cap})} \hat{\mathbf{N}} \cdot \overline{\rho_{\text{m}} \mathbf{v}}\, dS + \int_s \hat{\mathbf{N}} \cdot \overline{\eta_{\text{m}} \mathbf{v}}\, \mathbf{ds}. \tag{45}$$

The integral forms may be written in terms of the auxiliary vectors. In Eq. (40), **E** and $\bar{Q}$ are replaced by **D** and **Q**, respectively; Q is obtained from Eq. (44) by replacing $\bar{\rho}$ and $\bar{\eta}$ by ρ and η. In Eq. (42), **B** and $\bar{I}$ are replaced by **H** and I, respectively; I is obtained from Eq. (45) by replacing $\overline{\rho_{\text{m}} \mathbf{v}}$ by **J** and $\overline{\eta_{\text{m}} \mathbf{v}}$ by **K**.

E. Field Equations in Simple Media

Maxwell's equations in their general forms together with the definitions of the essential densities involve four volume densities: ρ, **P**, **J**, and **M**. In many materials, these densities are induced by an externally maintained electromagnetic field. In simple media or linear media, there is a linear relation between the density functions on the one hand and the exciting field on the other. In such media,

$$\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P} = (1 + \chi_{\text{e}}) \varepsilon_0 \mathbf{E}, \tag{46}$$

$$\mathbf{H} = \mu_0^{-1} \mathbf{B} - \mathbf{M} = (1 + \chi_{\text{m}}) \mu_0^{-1} \mathbf{B}, \tag{47}$$

$$\mathbf{J}_f = \sigma \mathbf{E}, \tag{48}$$

where χ_{e} is the electric susceptibility, χ_{m} the magnetic susceptibility, and σ the conductivity. The quantities $(1 + \chi_{\text{e}})$ and $(1 + \chi_{\text{m}})$; are each represented by a symbol standing for the properties of linear polarizability or linear magnetizability.Hence $\varepsilon_{\text{r}} = (1 + \chi_{\text{e}})$ and $\mu_{\text{r}} = (1 + \chi_{\text{m}})$; ε_{r} is the relative dielectric constant or relative permittivity of a linearly polarizable medium and μ_{r} is the relative permeability of a linearly magnetizable medium. Then $\varepsilon = \varepsilon_0 \varepsilon_{\text{r}}$ and $\mu = \mu_0 \mu_{\text{r}}$; ε is the dielectric constant or permittivity and μ is the permeability. The auxiliary vectors and the polarization and magnetization vectors may be written in terms of these symbols.

The field equations in simple media can now be written as

$$\varepsilon \nabla \cdot \mathbf{E} = \rho_{\text{f}}, \tag{49}$$

$$\nabla \times \mathbf{E} = -\partial \mathbf{B} / \partial t, \tag{50}$$

$$\mu^{-1} \nabla \times \mathbf{B} = \sigma \mathbf{E} + \varepsilon\, \partial \mathbf{E} / \partial t, \tag{51}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{52}$$

where the volume density of free charge ρ_{f} is the only density function appearing explicitly. Equations (49)–(52) are valid in linearly polarizing, magnetizing, and conducting media.

The boundary conditions are also expressed solely in terms of free charge densities as follows:

$$\varepsilon_1 \hat{\mathbf{n}}_1 \cdot \mathbf{E}_1 + \varepsilon_2 \hat{\mathbf{n}}_2 \cdot \mathbf{E}_2 = -(\eta_{1\mathrm{f}} + \eta_{2\mathrm{f}}), \quad (53)$$

$$\hat{\mathbf{n}}_1 \times \mathbf{E}_1 + \hat{\mathbf{n}}_2 \times \mathbf{E}_2 = 0, \quad (54)$$

$$\mu_1^{-1}(\hat{\mathbf{n}}_1 \times \mathbf{B}_1) + \mu_2^{-1}(\hat{\mathbf{n}}_2 \times \mathbf{B}_2) = -(\mathbf{K}_{1\mathrm{f}} + \mathbf{K}_{2\mathrm{f}}), \quad (55)$$

$$\hat{\mathbf{n}}_1 \cdot \mathbf{B}_1 + \hat{\mathbf{n}}_2 \cdot \mathbf{B}_2 = 0. \quad (56)$$

The field equations and the boundary conditions may be specialized for different materials. For good conductors the conductivity is large, and it may be assumed that the free charge is on the surface; the volume density of charge is therefore approximately zero. Equations (49) and (51) now become

$$\nabla \cdot \mathbf{E} = 0 \quad (57)$$

and

$$\nabla \times \mathbf{B} = \mu\sigma\mathbf{E}. \quad (58)$$

At the boundary between two good conductors, numbered 1 and 2, with the assumption that $\mathbf{K}_{1f}$ and $\mathbf{K}_{2f}$ are each zero, Eq. (55) becomes

$$\mu_1^{-1}(\hat{\mathbf{n}}_1 \times \mathbf{B}_1) + \mu_2^{-1}(\hat{\mathbf{n}}_2 \times \mathbf{B}_2) = 0. \quad (59)$$

All the other equations are unchanged in the interior as well as on the surface.

For nonconductors, there is no free charge distribution on the surface or in the interior of the material and the conductivity is very small. Thus, the field equations (49) and (51) for the interior become

$$\nabla \cdot \mathbf{E} = 0 \quad (60)$$

and

$$\nabla \times \mathbf{B} = \mu\varepsilon\, \partial\mathbf{E}/\partial t. \quad (61)$$

At boundaries between two nonconductors, numbered 1 and 2, there are no free charges or currents. Hence, the right-hand sides of Eqs. (53) and (55) vanish. All the other equations remain unchanged. When there is a boundary between a nonconductor and a conductor, Eqs. (53)–(56) may be suitably adapted.

The first-order partial differential equations can be converted to second-order equations. Hence,

$$\nabla \times \nabla \times \mathbf{E} + \mu\sigma \frac{\partial \mathbf{E}}{\partial t} + \mu\varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} = 0. \quad (62)$$

The second-order equation for the magnetic field is obtained from Eq. (62) by replacing **E** by **B**.

The field equations have been formulated for arbitrary time dependence in the foregoing. For most practical applications, sinusoidal signals are used. Signals that are in the form of pulses can also be decomposed into sinusoidal components with the use of Fourier analysis. Therefore, Maxwell's equations are obtained assuming a periodic (or harmonic) time dependence.

The equations for the interior are

$$\varepsilon_0 \nabla \cdot \mathbf{E} = \bar{\rho}, \quad (63)$$

$$\nabla \times \mathbf{E} = -j\omega\mathbf{B}, \quad (64)$$

$$\nabla \times \mathbf{B} = \overline{\rho_{\mathrm{m}}\mathbf{v}} + j\omega\varepsilon_0\mathbf{E}, \quad (65)$$

$$\nabla \cdot \mathbf{B} = 0. \quad (66)$$

The boundary conditions are unchanged and are given by Eqs. (53)–(56). The definitions of $\bar{\rho}$, $\bar{\eta}$, and $\overline{\eta_{\mathrm{m}}\mathbf{v}}$ are unchanged. The volume current density function becomes

$$\overline{\rho_{\mathrm{m}}\mathbf{v}} = \mathbf{J} + \nabla \times \mathbf{M} + j\omega\mathbf{P}. \quad (67)$$

The equations of continuity are

$$\nabla \cdot \overline{\rho_{\mathrm{m}}\mathbf{v}} + j\omega\rho = 0 \quad (68)$$

and

$$\nabla \cdot (\overline{\eta_{\mathrm{m}}\mathbf{v}_1} + \overline{\eta_{\mathrm{m}}\mathbf{v}_2}) + j\omega(\overline{n_1} + \overline{\eta_2}) - (\hat{\mathbf{n}}_1 \cdot \overline{\rho_{\mathrm{m}}\mathbf{v}_1} + \hat{\mathbf{n}}_2 \cdot \overline{\rho_{\mathrm{m}}\mathbf{v}_2}) = 0. \quad (69)$$

The field vectors and the density functions are now of the form

$$\mathbf{E}(r, t) = \mathrm{Re}[\mathbf{E}(r, \omega)e^{j\omega t}]. \quad (70)$$

The field equations and boundary conditions for simple media with periodic time dependence are easily obtained from the above relations by expressing **J** and **P** in terms **E** and **M** in terms of **B** for nonconductors and conductors.

There are certain generalized coefficients for simple media that are of great importance in electromagnetics. The complex quantity k is known as the wavenumber and is defined as

$$k = \omega(\mu\varepsilon)^{1/2}, \quad (71)$$

where ε is a complex permittivity defined as

$$\varepsilon = (\varepsilon - j\sigma/\omega). \quad (72)$$

When the medium is air, $\varepsilon = \varepsilon_0$, $\mu = \mu_0$, and k becomes the free-space wavenumber $k_0 = \omega/c$, where c is the velocity of light; k is dimensionally a reciprocal length measured in reciprocal meters. The complex wavenumber defined in Eq. (71) may be written as $k = \beta - j\alpha$, where β is the real phase constant and α is the real attenuation constant. Corresponding to the velocity in air, a phase velocity in the medium is defined as $v = \omega/\beta$. The loss tangent p is defined as

$$p = \sigma/\omega\varepsilon, \quad (73)$$

so that

$$k = k_{\mathrm{e}}(1 - jp), \quad (74)$$

where $k_{\mathrm{e}} = \omega(\mu\varepsilon)^{1/2} = k_0\varepsilon_r^{1/2}$. The characteristic phase velocity v and characteristic impedance ζ of a simple

medium may be defined by analogy with air with ε and μ replacing ε_0 and μ_0.

For poor conductors, the loss tangent is very small, $p^2 \ll 1$ and $\beta = k_e = \omega/v$. The attenuation is also very small and is given by $\alpha = \sigma\zeta/2$. For good conductors, $p \gg 1$, $\varepsilon_r = 0$ and $\beta = \alpha = (\omega\sigma\mu/2)^{1/2}$. The reciprocal of α is dimensionally a length called the skin depth or skin thickness for a good conductor. It is $d_s = 1/\alpha = (2/\omega\mu\sigma)^{1/2}$. It is defined as the depth at which the electric field reduces to $1/e$ of its value at the surface of the conductor, where e is the base of the Napierian logarithm.

F. Scalar and Vector Potential Functions

In the solution of electromagnetic field problems, it is often easier to use auxiliary functions defined in terms of the field vectors such as potential functions. Two such functions are the scalar potential function ϕ and the vector potential function $\mathbf{A}$. The four first-order field equations can be transformed into two second-order equations in the potential functions that can be integrated. The magnetic vector $\mathbf{B}$ has a zero divergence and it is said to be a solenoidal vector. Using the vector identity $\nabla \cdot \nabla \times \mathbf{C} = 0$, one can derive, $\mathbf{B}$ from a vector potential function $\mathbf{A}$,

$$\nabla \times \mathbf{A} = \mathbf{B}. \tag{75}$$

The vector point function is defined incompletely in Eq. (75), as its divergence is not known. This function is known as the *magnetic vector potential*.

The define a scalar potential it is necessary to find a vector with vanishing curl. From the symmetry of electric and magnetic quantities, the second field equation should be used for this purpose, since this is the electric analogue of the magnetic fourth equation. The second equation in its most general form is given in Eq. (17). With the substitution of Eq. (75), this becomes

$$\nabla \times (\mathbf{E} + \partial\mathbf{A}/\partial t) = 0. \tag{76}$$

The vector $(\mathbf{E} + \partial\mathbf{A}/\partial t)$ is a potential vector because its curl vanishes. It can be derived from a scalar potential ϕ defined by

$$-\nabla\phi = \mathbf{E} + \partial\mathbf{A}/\partial t. \tag{77}$$

This is also a fundamentally important relation. The scalar potential defined by Eq. (77) is called the *electric scalar potential*. It is seen that if the scalar and vector potentials ϕ and $\mathbf{A}$ are known, the electromagnetic vectors $\mathbf{E}$ and $\mathbf{B}$ may be determined directly from them.

With the scalar and vector potentials defined, the next step is to eliminate $\mathbf{E}$ and $\mathbf{B}$ from the field equations. It is to be noted that the scalar potential ϕ has been defined completely, but the definition of $\mathbf{A}$ will not be complete until its divergence is defined. The potential functions have been defined in terms of two of the field equations. They must still satisfy the other two. The divergence of $\mathbf{A}$ is now defined as follows:

$$\nabla \cdot \mathbf{A} + \varepsilon_0\mu_0\, \partial\phi/\partial t = 0. \tag{78}$$

This is known as the Lorentz condition. With Eq. (78) and the relations between the field vectors and the potential functions, the following d'Alembert equations governing the potential functions ϕ and $\mathbf{A}$ are obtained:

$$\nabla^2\phi - \mu_0\varepsilon_0\, \partial^2\phi/\partial t^2 = -\bar{\rho}/\varepsilon_0. \tag{79}$$

and

$$\nabla^2\mathbf{A} - \mu_0\varepsilon_0\, \partial^2\mathbf{A}/\partial t^2 = -\mu_0\overline{\rho_m \mathbf{v}}. \tag{80}$$

In the stationary state, the second terms in Eqs. (79) and (80) vanish and each reduces to Poisson's equation. In regions where there are no densities present, Poisson's equation reduces to Laplace's equation. In simple media, the equations may be obtained directly by writing ε for ε_0, μ for μ_0, ρ_f for $\bar{\rho}$, and J_f for $\overline{\rho_m \mathbf{v}}$. The potential equations and the Lorentz condition may be written for periodic time dependence by the substitutions $\partial/\partial t = j\omega$ and $\partial^2/\partial t^2 = -\omega^2$.

The solutions to the potential equations with periodic time dependence may be written in different forms, but it is desirable to choose those that satisfy the Lorentz condition. A particularly useful form of the solution is a particular integral called the Helmholtz integral. These integrals for the scalar and vector potentials are

$$\phi = \frac{1}{4\pi\varepsilon_0}\left[\int_\tau \frac{\bar{\rho}'}{R}\, e^{-jk_0R}\, d\tau' + \int_\Sigma \frac{\bar{\eta}'}{R}\, e^{-jk_0R}\, d\sigma'\right], \tag{81}$$

$$\mathbf{A} = \frac{\mu_0}{4\pi}\left[\int_\tau \frac{\overline{\rho_m \mathbf{v}'}}{R}\, e^{-jk_0R}\, d\tau' + \int_\Sigma \frac{\overline{\eta_m \mathbf{v}'}}{R}\, e^{-jk_0R}\, d\sigma'\right], \tag{82}$$

where $d\tau'$ represents an element of volume, and $d\sigma'$ represents an element of surface.The integration is carried out over all regions and surfaces where the density functions are nonzero and are defined. R is the distance from a point P where ϕ and $\mathbf{A}$ are to be determined to a variable point of integration P'. In Cartesian coordinates, R is the distance between $P(x, y, z)$ and $P'(x', y', z')$ and given by $R = [(x - x')^2 + (y - y')^2 + (z - z')^2]^{1/2}$. Equations (81) and (82) represent the solutions to the D'Alembert equations for the scalar and vector potentials for a periodic time dependence outside regions where the charge and current densities exist. The solutions in the interior of regions where charge and current densities exist are obtained by solving the appropriate boundary value problem. It may be noted that $k_0 = \omega(\varepsilon_0\mu_0)^{1/2}$.

G. Electromagnetic Force and Energy

The concepts of force and energy are very important in any discussion of electromagnetics. The electromagnetic force and energy may be expressed in terms of the field vectors as well as the potential functions. The electromagnetic force acting on a body is given by

$$\mathbf{F} = q\mathbf{E} + q_{\mathrm{m}}\mathbf{v} \times \mathbf{B}, \tag{83}$$

with

$$q = \int_{\tau} \bar{\rho}\, d\tau + \int_{\Sigma} \bar{\eta}\, d\sigma, \tag{84}$$

$$q_{\mathrm{m}}\mathbf{v} = \int_{\tau} \overline{\rho_{\mathrm{m}}\mathbf{v}}\, d\tau + \int_{\Sigma} \overline{\eta_{\mathrm{m}}\mathbf{v}}\, d\sigma. \tag{85}$$

If all the charges are in motion and none is stationary (as, for example, in an electron stream), q and $q_{\mathrm{m}}\mathbf{v}$ coincide and Eq. (83) reduces to

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{86}$$

A vector $\mathbf{S}$ called the Poynting vector is defined as

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}. \tag{87}$$

It has the units of volt-ampere per square meter or watt per square meter. The total outward normal flux of the Poynting vector $\mathbf{T}$ is given by

$$\mathbf{T} = \int_{\Sigma} (\hat{n} \cdot S)\, d\sigma, \tag{88}$$

where Σ represents the surface enclosing the volume τ. The electric energy density is equal to $\mathbf{D} \cdot \mathbf{E}$, and the magnetic energy density is equal to $\mathbf{H} \cdot \mathbf{B}$. Therefore, the total electromagnetic energy stored in the volume τ is

$$U = U_{\mathrm{E}} + U_{\mathrm{M}} = \int_{\tau} \left(\tfrac{1}{2}\mu^{-1}B^2 + \tfrac{1}{2}\varepsilon E^2\right) d\tau. \tag{89}$$

The principle of conservation of energy requires that the net energy lost must be equal to the net energy gained. Since power is the time rate of change of energy, the principle of energy conservation can be written as the principle of power conservation. The Poynting theorem gives the power conservation equation for simple media,

$$T = -\partial U/\partial t - \int_{\tau} \mathbf{J}_{\mathrm{f}} \cdot \mathbf{E}\, d\tau \tag{90}$$

T measures the time rate of the increase of energy associated with all the regions outside the volume τ; that is, it represents the flow of power across the surface Σ enclosing the volume τ. The element U is defined in Eq. (89) and the first term on the right of Eq. (90) defines the time rate of decrease of the total electromagnetic energy in the volume τ; the second term on the right represents the time rate of heat dissipated in the volume τ, and it is associated with moving free charges. It is observed that as the energy (or power) outside τ increases, the energy (or power) inside τ decreases, thus conserving energy. Equation (90) is obtained from Maxwell's equations. The freespace Poynting theorem is written by substituting ε_0 for ε and μ_0 for μ in Eq. (90).

For harmonic time dependence, a complex Poynting vector may be defined as

$$\mathbf{S}^* = \tfrac{1}{2}(\mathbf{E} \times \mathbf{H}). \tag{91}$$

A real power equation is now obtained:

$$\int_{\Sigma} \mathrm{Re}(\mathbf{S}^*)\, d\sigma = \frac{1}{2}\int_{\tau} \frac{J_{\mathrm{f}}^2}{\sigma}\, d\tau, \tag{92}$$

where $\mathrm{Re}(\mathbf{S}^*)$ is the time-average power density leaving the unit area of the surface Σ. The time-average electric and magnetic energy functions may be defined as $\langle U_{\mathrm{E}} \rangle = U_{\mathrm{E}}/2$ and $\langle U_{\mathrm{M}} \rangle = U_{\mathrm{M}}/2$. The total time-average electromagnetic energy is the sum of the two quantities. It remains a constant due to energy conservation, and hence the time-average electric energy is equal to the time-average magnetic energy, and the total energy may be written as

$$\langle U(t) \rangle = 2\langle U_{\mathrm{E}}(t) \rangle = 2\langle U_{\mathrm{M}}(t) \rangle. \tag{93}$$

III. ELECTROMAGNETIC WAVES

The Helmholtz integrals that define the potential field in terms of the essential densities of charge and current were seen in Eqs. (81) and (82). Since the relations between the electric and magnetic vectors and the potential functions are known, explicit formulas for $\mathbf{E}$ and $\mathbf{B}$ may be obtained. These are the integral solutions of Maxwell's equations. The general integrals for the electromagnetic field with periodic time dependence are

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\int_{\tau} e^{-jk_0R}\left[\frac{\hat{\mathbf{R}}\bar{\rho}'}{R^2} + \left(\bar{\rho}'\hat{\mathbf{R}} - \frac{\overline{\rho_{\mathrm{m}}\mathbf{v}'}}{C}\right)\frac{jk_0}{R}\right] d\tau' + \mathbf{S}_{\mathrm{E}}, \tag{94}$$

$$\mathbf{B} = \frac{\mu_0}{4\pi}\int_{\tau} e^{-jk_0R}(\hat{\mathbf{R}} \times \overline{\rho_{\mathrm{m}}\mathbf{v}'})\left(\frac{1}{R^2} + \frac{jk_0}{R}\right) d\tau' + \mathbf{S}_{\mathrm{B}}, \tag{95}$$

where $\mathbf{S}_{\mathrm{E}}$ and $\mathbf{S}_{\mathrm{B}}$ are the surface integrals that are obtained from the volume integrals in each case by writing η for ρ and Σ for τ, $\hat{\mathbf{R}}$ is a unit vector from the variable point P' to the fixed point P where the field is to be calcualted, and R is the distance between P' and P.

If the electromagnetic field at points in space is due to currents and charges on a cylindrical conductor (such as a dipole antenna) with a small cross section and its axis

along the z axis of a coordinate system, the integrals in Eqs. (94) and (95) may be further simplified. The conductor is assumed to extend from $z=-h$ to $z=+h$ along the z axis and to have a radius a. The expressions for **E** and **B** are obtained from Eqs. (94) and (95) by replacing the volume integral by a line integral between the limits $z=-h$ and $z=+h$; $\bar{\rho}'$ becomes q', $\overline{\rho_m \mathbf{v}'}$ becomes $\hat{\mathbf{z}} I_z$, and $d\tau'$ becomes dz'. There is no surface integral and

$$q' = \int_S \bar{\rho}'\, dS' + \int_s \bar{\eta}'\, ds', \tag{96}$$

$$I_z' = \int_S \overline{\rho_m \mathbf{v}_z'}\, dS' + \int_s \overline{\eta_m \mathbf{v}_z'}\, ds'. \tag{97}$$

In Eqs. (96) and (97), the first integral is taken over the cross section $S=\pi a^2$ and the second integral is taken over the circumference $s=2\pi a$; q' and I_Z' are the amplitudes of the total charge per unit length and the total axial conduction current. For a thin conductor, the contributions due to radial and azimuthal current are negligible. The instantaneous values of **E** and **B** are obtained by introducing the time factor $e^{j\omega t}$ and taking the real part. The charge is expressed in terms of the current using the equation of continuity:

$$dI_z'/dz' + j\omega q' = 0. \tag{98}$$

The vector **E** has a component along R and a second component parallel to $\hat{z}$ and hence parallel to the current in the conductor; **B** is every where perpendicular to both R and z and hence is directed along tangents to circles drawn with the conductor as the common axis. Therefore, **E** and **B** are everywhere mutually perpendicular, and the equiphase surfaces of **E** and **B** due to the periodically varying charge and current *in an element* dz' of the conductor are spherical shells.

The general electromagnetic field can be expressed as the sum of components: the induction field and the radiation field. The induction (or near-zone) field dominates in the immediate vicinity of the volume on and in which charge and current densities exist. The radiation (or far-zone) field is dominant at great distances from the source distributions. The near zone is characterized by the condition $k_0 R \ll 1$, and only the terms with $1/R^2$ are significant in the volume and surface integrals in Eqs. (94) and (95). In the stationary state, all the densities are constant in time, and the expressions for the electrostatic and magnetostatic field are obtained from Eqs. (94) and (95). At $\omega=0$ and at low frequencies, the induction field is dominant and the radiation field is negligibly small. In the special case where R is larger than the largest dimension in the volume τ, the field vector is given by

$$\mathbf{E} = \hat{\mathbf{R}}\, q \big/ 4\pi\varepsilon_0 R^2 = \mathbf{F} Q, \tag{99}$$

where **F** is the electrostatic force and Eq. (99) is a statement of Coulomb's law, a fundamental postulate of electromagnetism. The expression for the near-zone **B** vector is called the Biot–Savart law.

The radiation or far-zone field is defined by $k_0 R \gg 1$. When this condition is satisfied, the radiation field is dominant, and the terms involving $1/R$ in Eqs. (94) and (95) define the **E** and **B** vectors. The $1/R^2$ terms are vansihingly small. As before, the fields may be specialized for the case of a cylindrical conductor that is thin.

The electromagnetic field in the far zone due to periodically varying charges and currents in an element of length dz' may be described in terms of a wave picture with spherically equiphase surfaces expanding with the constant velocity c in free space. The magnetic waves are transverse; the electric waves are both transverse and longitudinal.

Spherical and plane electromagnetic waves will now be described. Since the conductor of length $2h$ producing the field is at a very large distance from the point of observation P where the radiation field is to be determined, the inequality $R \gg h$ is satisfied. Here, R is the distance from any point on the conductor to P. It follows that $R_0 \gg h$ is also satisfied, where R_0 is the distance from the origin located at the center of the conductor to P, as shown in Fig. 4. Then the field vectors may be written as

$$\mathbf{E} = \frac{jk_0}{4\pi\varepsilon_0}\frac{e^{-jk_0R_0}}{R_0}\int_{-h}^{+h}\left(\hat{\mathbf{R}}_0 q' - \frac{\hat{\mathbf{z}} I_z'}{c}\right)e^{-jk_0R_0\cos\theta}\, dz' \tag{100}$$

and

$$\mathbf{B} = \frac{j\mu_0 k_0}{4\pi}\frac{e^{-jk_0R_0}}{R_0}\int_{-h}^{+h}(\hat{\mathbf{R}}_0 \times \hat{\mathbf{z}}) I_z'\, e^{-jk_0R_0\cos\theta}\, dz'. \tag{101}$$

It may be shown that

$$\mathbf{E} = c(\mathbf{B} \times \hat{\mathbf{R}}_0). \tag{102}$$

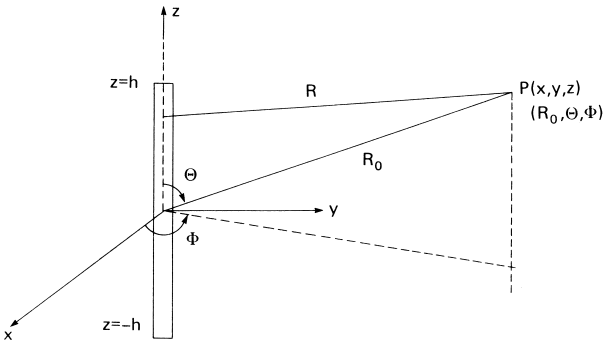


FIGURE 4 Antenna in space.

In the system of spherical coordinates (R_0, Θ, Φ) the **E** vector has a Θ component and the **B** vector has a Φ component. All other components vanish. The field in the radiation zone is of the form

$$E_\Theta = f(h, \Theta)\frac{e^{-jk_0R_0}}{R_0} = F(\Theta, R_0) = cB_\Phi, \quad (103)$$

with

$$f(h, \Theta) = \frac{j\omega\mu_0}{4\pi}\int_{-h}^{+h} I'_z\, e^{jk_0z'\cos\theta}\sin\Theta\, dz'. \quad (104)$$

The function F describes transverse spherical waves expanding radially outward with a constant velocity c. The spherical equiphase surfaces are separated by a constant distance, which is the wavelength $\lambda_0 = 2\pi/k_0$.

In most problems involving the far-zone field of a radiating structure, only a small part of space is involved. The function F can then be approximated by

$$Fe^{j\omega t} = Ke^{j(\omega t - k_0 s)}. \quad (105)$$

Equation (105) characterizes the approximate distribution of **E** and **B** at a large distance from the radiating structure within a volume V whose dimensions are small compared with R. It is readily interpreted in terms of a simple wave picture. The element s may be written in terms of the Cartesian coordinates as

$$s = xl_0 + ym_0 + zn_0, \quad (106)$$

where l_0, m_0 and n_0 are the direction cosines, defined as the cosines of the angles between each of the coordinate axes and R_0. Equation (106) defines a plane at right angles to s and at a distance s from the origin. Then Eq. (105) may be described using a picture of plane equiphase surfaces at right angles to s and traveling along s with a constant velocity c. The arcs of radially expanding spherical equiphase surfaces defined by Eq. (104) are assumed to be approximately plane and of constant amplitude provided the distance to the volume from the source is sufficiently great and the solid angle subtended by it is small. Hence, the spherical electromagnetic wave may be approximated by the plane electromagnetic wave.

In problems involving the field in the radiation zone, it is convenient to use rectangular coordinates. The x axis coincides with the spherical coordinate R, the y axis is tangent to and in the direction of the coordinate Φ at P, and z axis is tangent to and in the direction $-\Theta$ at P; x, y, and z form a right-handed system, and the electromagnetic field at P in a small region surrounding it is given by

$$E_\Theta = -E_{z'}, \qquad B_\Phi = B_y. \quad (107)$$

In this case, the radiation zone field may be written as

$$-E_z(t) = cB_y(t) = Ke^{j(\omega t - k_0 x)}. \quad (108)$$

This expression is interpreted in terms of plane transverse waves normal to the x axis and traveling along it with a constant velocity c. The energy is transported in the x direction, and the Poynting vector is in the x direction.

The **E** vector is always parallel to the z axis and is said to be linearly polarized along the z axis. The **B** vector is linearly polarized along the y axis. If the orientation of the source of the radiation field, the antenna, is changed, the axes of polarization of **E** and **B** are changed. However, the electric and magnetic fields at any point in the radiation zone of a dipole antenna are always linearly polarized along mutually perpendicular axes. Furthermore, these axes lie in a plane at right angles to the line joining the point to the center of the antenna.

If there is more than one antenna, the field will still be in a plane at right angles to the line from the center of the antenna to the point of observation, but **E** and **B** will each have components in the z and y directions. If only one frequency is involved,

$$\mathbf{E}(t) = -(\hat{z}E_z + \hat{y}E_y)e^{j\omega t} \quad (109)$$

and

$$\mathbf{B}(t) = (\hat{y}B_y + \hat{z}B_z)e^{j\omega t}, \quad (110)$$

with

$$-E_z = cB_y = Ke^{-jk_0x}, \qquad -E_y = cB_z = Ne^{-jk_0x}, \quad (111)$$

where K and N are complex numbers given by

$$K = ae^{jg}; \qquad N = be^{jp}. \quad (112)$$

When $a \neq b$ and $d = p - g \neq 0$, the loci of the ends of the vectors $\mathbf{E}(t)$ and $\mathbf{B}(t)$ are ellipses. When $a = b$ and $\delta = \pm\pi/2$, the loci are circles, and when $\delta = \pi$, the fields are linearly polarized.

IV. APPLICATIONS OF ELECTROMAGNETICS

A. Transmission of Electromagnetic Waves

A generator producing an electric signal is the source of electromagnetic energy. This energy is transmitted from the generator in a wave motion in metallic or nonmetallic structures, which deliver the energy either to radiating structures such as antennas or to transformers at low frequencies. Up to the ultrahigh-frequency range, these guiding structures are transmission lines, which consist of two or more conductors placed in a specific configuration such as a coaxial one for two conductors. For higher frequencies, waveguides, strip lines, and microstrips are used. The electromagnetic field in these structures is obtained by solving the second-order partial differential equations

obtained from Maxwell's equations with the appropriate boundary conditions. For transmission lines, the field is of the transverse electromagnetic type, where the electric and magnetic fields are orthogonal to each other and transverse to the direction of propagation. The electromagnetic field for waveguides has components transverse to the direction of propagation as well as in the direction of propagation. The fields are of the transverse electric type (the component of the electric field in the direction of propagation vanishes) and of the transverse magnetic type (the component of the magnetic field in the direction of propagation vanishes). Approximate solutions of Maxwell's equations show that the field in microstrips is of the quasi-transverse electromagnetic type. At frequencies in the optical range, dielectric waveguides of very small cross section are used. The wave equation is solved subject to the boundary conditions for dielectrics. For such structures, hybrid modes are obtained with field components in all the coordinate directions.

B. Radiation of Electromagnetic Waves

Power from the generator is transmitted in a wave motion and delivered to the structure that radiates the energy into space. The radiating structure is called an antenna. The radiated field is determined by using the integral solutions for the second-order differential equations. It is usually more convenient to use the Helmholtz integrals for the scalar and vector potential functions and to determine the electromagnetic field from them. The current and charge distributions on the antenna have to be assumed in order to be able to use the integral solutions. The Helmholtz integrals may also be used to determine the current distribution on the antenna. The boundary conditions on the conducting surface of the antenna are used, and this yields an integral equation for the current that may be solved approximately. The electromagnetic field in the radiation zone depends on the structure of the antenna. In practical problems, the antenna system is designed to give the desired field.

C. Reception of Electromagnetic Waves

The radiated electromagnetic field is received by an antenna system consisting of one or more antennas. The properties of the antenna have to be known for it to be used efficiently as a receiver. An important theorem of electromagnetics is used to show that the properties of an antenna used as a receiver can be predicated from its performance as a transmitter. This is called the Rayleigh–Carson reciprocal theorem. It is derived from Maxwell's equations and subject only to the condition that the essential density of moving charge $\overline{\rho_{\mathrm{m}}\mathbf{v}}$ is linearly related to the electric field. It can be shown that this is true in all simply polarizing, magnetizing, and conducting media. The theorem may be expressed as follows: if a generator with an electromotive force (EMF) or driving potential difference of complex amplitude V^e between its terminals maintains a current of complex amplitude I through a load connected between any other pair of terminals in the same or in a coupled network, the current in the load is unaltered if load and generator are interchanged provided that the impedances connected between each pair of terminals are the same in both cases and the generator maintains the same EMF. Suppose that when $V_j^{e'}$ is applied at the terminals j, a current I_i' exists at terminals i, and when $V_i^{e''}$ is applied at terminals i, a current I_j'' exists at terminals j. The reciprocal theorem reduces to the form

$$I_j'' V_j^{e'} = I_i' V_i^{e''}. \tag{113}$$

The theorem may be further simplified if the same potential difference is applied in the one case across the terminals j as in the other case across the terminals i. When

$$V_j^{e'} = V_i^{e''}, \tag{114}$$

it follows that the reciprocal theorem becomes

$$I_j' = I_i' \tag{115}$$

When this relation is applied to an antenna, it can be shown that the directional properties of the antenna are the same for transmission and reception. Hence, the energy absorbed by a receiving antenna can be determined from its radiation zone field when used as a transmitting antenna.

D. Scattering and Diffraction of Electromagnetic Waves

Electromagnetic energy is transmitted from generators to antennas or other radiating systems in which oscillating currents are set up over a wide band of frequencies corresponding to the frequencies of the generator. These in turn induce similar currents in surrounding bodies and regions of matter. This interaction has been described earlier in terms of trains of electromagnetic waves traveling outward from the source. When they encounter obstacles, currents are induced in them, and this results in the generation of a secondary electromagnetic field known as the *scattered* or *reradiated field*. Where it penetrates the geometrical shadow, it is known as the *diffracted field*. The nature of the scattered and diffracted field depends on the electrical properties, shape, and orientation of the scattering obstacle relative to the incident field from the distant antenna. The actual calculation of the scattered field and the induced currents that generate it is carried out by solving

Maxwell's equations with the associated boundary conditions. A few of the problems are analytically solvable, among them being that of the scattering and diffraction by a conducting or totally absorbing (black) half-plane. This is one of the most important applications of electromagnetics in frequencies ranging from the radio to the optical range.

The problem is most easily solved by considering the reflection of plane electromagnetic waves from a perfectly conducting, infinite half-plane: An important theorem of electromagnetics is used to replace the half-plane by an equivalent source called the virtual source. The total field is then due to the actual source and the virtual source. The *theorem of images* determines the virtual source. If a conductor is placed above a perfectly conducting plane, the plane can be replaced by an identical image conductor arranged to be the exact geometrical image of the first conductor except in one respect. All the currents and charges, while the same in magnitude at image points in the image conductor, are opposite in direction and sign, respectively (Fig. 5). The resultant scalar and vector potentials at any point P are the sum of the potentials due to the actual conductor and its image. Since the conducting plane is at $z = 0$. the fields due to the actual and image conductor should vanish at $z = 0$. The theorem of images permits the substitution of a relatively simple problem involving a conductor and its image in space for a rather difficult one involving a single conductor over a perfectly conducting plane.

The virtual source for the scattering problem is obtained from the theorem of images, and the complete field is then the incident field together with the reflected field, which is the field of the virtual source. The scattered field may be calculated from Maxwell's equations and appropriate boundary conditions. When the geometry of the scattering obstacle changes, similar methods may be employed to determine the field. The scattering of the electromagnetic field form conducting cylinders and spheres can be solved analytically.

In all the applications of electromagnetics, experimental studies have to be carried out to verify theoretical conclusions. The experiment is often simplified if models of convenient size are used to simulate the system under study. The scale model may be much smaller or much larger than the actual system and of a size suitable for laboratory experiments. There is a relationship between the size of the scale model and the actual system. The determination of this relationship with the use of Maxwell's equations is called *electrodynamical similitude*. This theory of models provides the relationship between lengths, frequencies, conductivities, permittivities, and permeabilities in the system under study and the scale model.

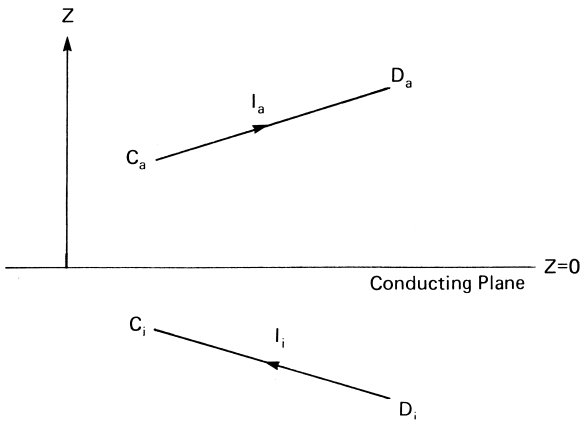


FIGURE 5 Conductor with image.

V. RECENT DEVELOPMENTS

The understanding of electromagnetics has been much enhanced by the use of computer simulation and optimization. Extensive software has become available for both personal and mainframe computers. Typical of such software are interactive graphics programs which show coordinate transformations, fields, and operators as well as electromagnetic wave motion. It is now possible to visualize the electromagnetic field and its behavior. Video cassettes have also become available which give a further appreciation of the electromagnetic field.

Computational methods for the solution of electromagnetic problems have also attained greater sophistication and these new numerical techniques have led to more accurate solutions. New computer-aided methods of numerical modeling are being developed, and the study of electromagnetics has taken on a new dimension.

SEE ALSO THE FOLLOWING ARTICLES

ELECTRODYNAMICS, QUANTUM • ELECTROMAGNETIC COMPATABILITY • FERROMAGNETISM • GEOMAGNETISM • MAGNETIC MATERIALS • RADIO PROPAGATION • SOLAR SYSTEM, MAGNETIC AND ELECTRIC FIELDS • WAVE PHENOMENA

BIBLIOGRAPHY

De Wolf, D. A. (2001). "Essentials of Electromagnetics for Engineering," Cambridge University Press, Cambridge.

Elliott, R. S. (1999). "Electromagnetics: History, Theory, and Application," IEEE, Piscataway, NJ.

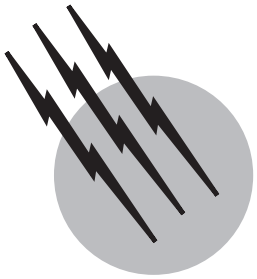
Hayt, W., and Buck, J. (2000). "Engineering Electromagnetics" 6th ed., McGraw-Hill, New York.

IEEE. (2000). "1999 International Conference on Computational Electromagnetics and Its Application," IEEE, Piscataway, NJ.

IEEE. (2000). "The Second Asia-Pacific Conference on Environmental Electromagnetics—CEEM 2000," IEEE, Piscataway, NJ.

Kraus, J. D., and Fleisch, D. (eds.). (1999). "Electromagnetics," McGraw-Hill, New York.

Peterson, A., Scott, R., and Mittra, R. (1998). "Computational Methods for Electromagnetics," Oxford University Press, Oxford.



Gravitational Wave Physics

Kostas D. Kokkotas
Aristotle University of Thessaloniki

I. Introduction
II. Theory of Gravitational Waves
III. Detection of Gravitational Radiation
IV. Astronomical Sources of Gravitational Waves

GLOSSARY

Black hole A region of spacetime in which there is such an immense concentration of material within a small volume that spacetime curves over on itself and the escape velocity from it exceeds the speed of light.
Neutron star A very compact "dead" star, whose degenerate material is composed almost entirely of neutrons.
Parsec (pc) The distance at which a body would have an annual parallax of 1 sec arc. It equals 3.26 light-years or 3.084×10^{13} km.
Pulsar A rotating neutron star, with an off-axis magnetic field, that emits regular pulses of radiation.
Supernova A stellar outburst during which a star (close to the end of its life) suddenly increases in brightness roughly 1 million times.
Virgo cluster A vast cluster of thousands of galaxies, of which 2500 are fairly bright. The average distance from earth is about 16 Mpc. So called because its center appears to lie in the constellation Virgo.

GRAVITATIONAL WAVES are propagating fluctuations of gravitational fields, that is, "ripples" in spacetime, generated mainly by moving massive bodies. These distortions of spacetime travel with the speed of light. Every body in the path of such a wave feels a tidal gravitational force that acts perpendicular to the wave's direction of propagation; these forces change the distance between given points, and the size of the change is proportional to the distance between the points. Gravitational waves can be detected by devices which measure the induced length changes. The frequencies and the amplitudes of the waves are related to the motion of the masses involved. Thus, the analysis of gravitational waveforms allows us to learn about their source and, if there are more than two detectors involved in observation, to estimate the distance and position of their source in the sky.

I. INTRODUCTION

Einstein first postulated the existence of gravitational waves in 1916 as a consequence of his theory of general relativity, but no direct detection of such waves has been made yet. The best evidence thus far for their existence is due to the work of 1993 Nobel laureates Joseph Taylor and Russell Hulse. They observed in 1974 two neutron

stars orbiting faster and faster around each other, exactly what would be expected if the binary neutron star was losing energy in the form of emitted gravitational waves. The predicted rate of orbital acceleration caused by gravitational radiation emission according to general relativity was verified observationally, with high precision.

Cosmic gravitational waves, upon arriving on earth, are much weaker than the corresponding electromagnetic waves. The reason is that strong gravitational waves are emitted by *very* massive compact sources undergoing *very* violent dynamics. These kinds of sources are not very common and so the corresponding gravitational waves come from large astronomical distances. On the other hand, the waves thus produced propagate essentially unscathed through space, without being scattered or absorbed from intervening matter.

A. Why Are Gravitational Waves Interesting?

Detection of gravitational waves is important for two reasons: First, their detection is expected to open up a new window for observational astronomy since the information they carry is very different from that carried by electromagnetic waves. This new window onto the universe will complement our view of the cosmos and will help us unveil the fabric of spacetime around black-holes, observe directly the formation of black holes or the merging of binary systems consisting of black holes or neutron stars, search for rapidly spinning neutron stars, dig deep into the very early moments of the origin of the universe, and look at the very center of the galaxies where supermassive black holes weighing millions of solar masses are hidden. These are only some of the great scientific discoveries scientists can expect to make during the early years of the 21st century. Second, detecting gravitational waves is important for our understanding of the fundamental laws of physics; the proof that gravitational waves exist will verify a fundamental 85-year-old prediction of general relativity. Also, by comparing the arrival times of light and gravitational waves from, e.g., supernovae, Einstein's prediction that light and gravitational waves travel at the same speed could be checked. Finally, we could verify that they have the polarization predicted by general relativity.

B. How Will We Detect Them?

Up to now, the only indication of the existence of gravitational waves is the indirect evidence that the orbital energy in the Hulse–Taylor binary pulsar is drained away at a rate consistent with the prediction of general relativity. The gravitational wave is a signal, the shape of which depends upon the changes in the gravitational field of its source. As mentioned earlier, any body in the path of the wave will feel an oscillating tidal gravitational force that acts in a plane perpendicular to the wave's direction of propagation. This means that a group of freely moving masses placed on a plane perpendicular to the direction of propagation of the wave will oscillate as long as the wave passes through them, and the distance between them will vary as a function of time, as in Fig. 1. Thus, the detection of gravitational waves can be accomplished by monitoring the tiny changes in the distance between freely moving test masses. These changes are extremely small; for example, when the Hulse–Taylor binary system finally merges, the strong gravitational wave signal that will be emitted will induce changes in the distance of two particles on earth that are 1 km apart much smaller than the diameter of the atomic nucleus! To measure such motions of macroscopic objects is a tremendous challenge for experimentalists. As early as the mid-1960s, Joseph Weber designed and constructed heavy metal bars, seismically isolated, to which a set of piezoelectric strain transducers were bonded in such a way that they could detect vibrations of the bar if it had been excited by a gravitational wave. Today, there are a number of such apparatuses operating around the world which have achieved unprecedented sensitivities, but they still are not sensitive enough to detect gravitational waves. Another form of gravitational wave detector that is more promising uses laser beams to measure the distance between two well-separated masses. Such devices are basically kilometer-sized laser interferometers consisting of three masses placed in an L-shaped configuration. The laser beams are reflected back and forth between the mirrors attached to the three masses, the mirrors lying several kilometers away from each other. A gravitational wave passing by will cause the lengths of the two arms to oscillate with time. When one arm contracts, the other expands, and this pattern alternates. The result is that the interference pattern of the two laser beams changes with time. With this technique, higher sensitivities could be achieved than are possible with the bar detectors. It is expected that laser interferometric detectors are the ones that will provide us with the first direct detection of gravitational waves.

II. THEORY OF GRAVITATIONAL WAVES

Newton's theory of gravity has enjoyed great success in describing many aspects of our every-day life and additionally explains most of the motions of celestial bodies in the universe. General relativity corrected Newton's theory and is recognized as one of the most ingenious creations of the human mind. The laws of general relativity, though, in the case of slowly moving bodies and weak gravitational

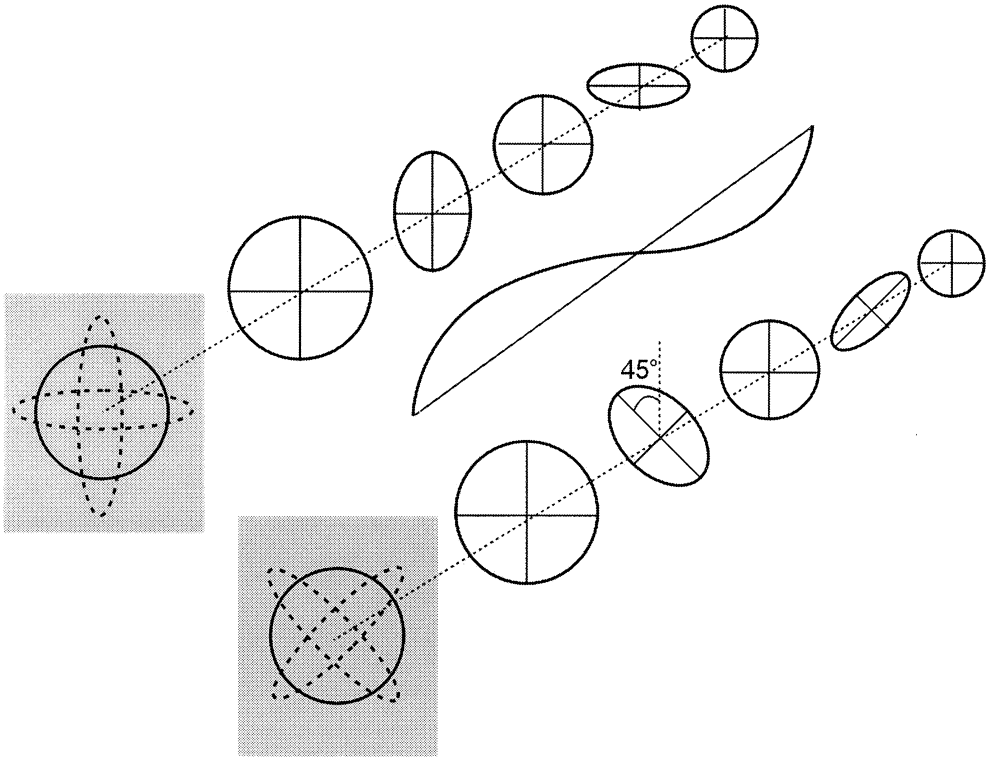


FIGURE 1 The effects of a gravitational wave traveling perpendicular the plane of a circular ring of particles, sketched as a series of snapshots. The deformations due the two polarizations are shown.

fields reduce to the standard laws of Newtonian theory. Nevertheless, general relativity is conceptually different from Newton's theory as it introduces the notion of spacetime and its geometry. One of the basic differences of the two theories concerns the speed of propagation of any change in a gravitational field. As the apple falls from the tree, we have a rearrangement of the distribution of mass of the earth, the gravitational field changes, and a distant observer with a high-precision instrument will detect this change. According to Newton, the changes of the field are instantaneous, i.e., they propagate with infinite speed; if this were true, however, the principle of causality would break down. No information can travel faster than the speed of light. In Einstein's theory there is no such ambiguity; the information of the varying gravitational field propagates with finite speed, the speed of light, as a ripple in the fabric of spacetime. These are the gravitational waves. The existence of gravitational waves is an immediate consequence of any relativistic theory of gravity. However, the strength and the form of the waves depend on the details of the gravitational theory. This means that the detection of gravitational waves will also serve as a test of basic gravitational theory.

The fundamental geometrical framework of relativistic metric theories of gravity is spacetime, which mathematically can be described as a four-dimensional manifold whose points are called *events*. Every event is labeled by four coordinates x^μ ($\mu = 0, 1, 2, 3$); the three coordinates x^i ($i = 1, 2, 3$) give the spatial position of the event, while x^0 is related to the coordinate time t ($x^0 = ct$, where c is the speed of light, which unless otherwise stated will be set equal to 1). The choice of the coordinate system is quite arbitrary and coordinate transformations of the form $\tilde{x}^\mu = f^\mu(x^\lambda)$ are allowed. The motion of a test particle is described by a curve in spacetime. The distance ds between two neighboring events, one with coordinates x^μ and the other with coordinates $x^\mu + dx^\mu$, can be expressed as a function of the coordinates via a symmetric tensor $g_{\mu\nu}(x^\lambda) = g_{\nu\mu}(x^\lambda)$, i.e.,

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu. \tag{1}$$

This is a generalization of the standard measure of distance between two points in Euclidean space. For the Minkowski spacetime (the spacetime of special relativity), $g_{\mu\nu} \equiv n_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$. The symmetric tensor $g_{\mu\nu}$ is called the *metric tensor* or simply the metric of the spacetime. In general relativity the gravitational field is described by the metric tensor alone, but in many other theories one or more supplementary fields may be needed as well. In what follows, we will consider only the general

relativistic description of gravitational fields since most of the alternative theories fail to pass the experimental tests.

The information about the degree of curvature (i.e., the deviation from flatness) of a spacetime is encoded in the metric of the spacetime. According to general relativity, any distribution of mass bends the spacetime fabric, and the Riemann tensor $R_{\kappa\lambda\mu\nu}$ (which is a function of the metric tensor $g_{\mu\nu}$ and of its first and second derivatives) is a measure of the spacetime curvature. The Riemann tensor has 20 independent components. When it vanishes, the corresponding spacetime is flat.

In the following presentation, we will consider mass distributions, which we will describe by the stress-energy tensor $T^{\mu\nu}(x^\lambda)$. For a perfect fluid (a fluid or gas with isotropic pressure but without viscosity or shear stresses) the stress-energy tensor is given by the following expression:

$$T^{\mu\nu}(x^\lambda) = (\rho + p)u^\mu u^\nu + pg^{\mu\nu}, \tag{2}$$

where $p(x^\lambda)$ is the local pressure, $\rho(x^\lambda)$ is the local energy density, and $u^\mu(x^\lambda)$ is the four-velocity of the infinitesimal fluid element characterized by the event x^λ.

Einstein's gravitational field equations connect the curvature tensor (see below) and the stress-energy tensor through the fundamental relation

$$G_{\mu\nu} \equiv R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}R = kT_{\mu\nu}. \tag{3}$$

This means that the gravitational field, which is directly connected to the geometry of spacetime, is related to the distribution of matter and radiation in the universe. By solving the field equations, both the gravitational field (the $g_{\mu\nu}$) and the motion of matter is determined. $R_{\mu\nu}$ is the so-called Ricci tensor and comes from a contraction of the Riemann tensor ($R_{\mu\nu} = g^{\sigma\rho}R_{\sigma\mu\rho\nu}$), R is the scalar curvature ($R = g^{\rho\sigma}R_{\rho\sigma}$), while $G_{\mu\nu}$ is the so-called Einstein tensor, $k = 8\pi G/c^4$ is the coupling constant of the theory, and G is the gravitational constant, which, unless otherwise stated, will be considered equal to 1. The vanishing of the Ricci tensor corresponds to a spacetime free of any matter distribution. However, this does not imply that the Riemann tensor is zero. As a consequence, in the empty space far from any matter distribution, the Ricci tensor will vanish, while the Riemann tensor can be nonzero; this means that the effects of a propagating gravitational wave in an empty spacetime will be described via the Riemann tensor.

A. Linearized Theory

Now us assume that an observer is far away from a given static matter distribution, and the spacetime in which he or she lives is described by a metric $g_{\mu\nu}$. Any change in the matter distribution, i.e., in $T_{\mu\nu}$, will induce a change in the gravitational field, which will be recorded as a change in metric. The new metric will be

$$\tilde{g}_{\mu\nu} = g_{\mu\nu} + h_{\mu\nu}, \tag{4}$$

where $h_{\mu\nu}$ is a tensor describing the variations induced in the spacetime metric. As we will describe analytically later, this new tensor describes the propagation of ripples in spacetime curvature, i.e., the *gravitational waves*. In order to calculate the new tensor $h_{\mu\nu}$ we have to solve Einstein's equations for the varying matter distribution. This is not an easy task in general. However, there is a convenient, yet powerful way to proceed, namely to assume that $h_{\mu\nu}$ is small ($|h_{\mu\nu}| \ll 1$), so that we need only keep terms linear in $h_{\mu\nu}$ in our calculations. In making this approximation we are effectively assuming that the disturbances produced in spacetime are not huge. This *linearization approach* has proved extremely useful for calculations, and, for weak fields at least, gives accurate results for the generation of the waves and for their propagation.

The first attempt to prove that in general relativity gravitational perturbations propagate as waves with the speed of light is due to Einstein himself. Shortly after the formulation of his theory—the year after—he proved that by assuming linearized perturbations around a flat metric, i.e., $g_{\mu\nu} = n_{\mu\nu}$, then the tensor

$$\tilde{h}_{\mu\nu} \equiv h_{\mu\nu} - \frac{1}{2}n_{\mu\nu}h^\alpha_\alpha \tag{5}$$

is governed by a wave equation which admits plane wave solutions similar to the ones of electromagnetism; here $h_{\mu\nu}$ is the metric perturbation and $\tilde{h}_{\mu\nu}$ is the gravitational field (or the "trace reverse" of $h_{\mu\nu}$). Then the *linear field equations in vacuum* have the form

$$\left(-\frac{\partial^2}{\partial t^2} + \nabla^2\right)\tilde{h}^{\mu\nu} \equiv \partial_\lambda\partial^\lambda\tilde{h}^{\mu\nu} = 0 \tag{6}$$

($\partial_\mu k^\alpha \equiv \partial k^\alpha/\partial x^\mu$), which is the three-dimensional wave equation. To obtain the above simplified form, the condition $\partial_\mu\tilde{h}^{\mu\nu} = 0$, known as Hilbert's gauge condition (equivalent to the Lorentz gauge condition of electromagnetism), has been assumed. A gauge transformation is a suitable change of coordinates defined by

$$x'^\mu \equiv x^\mu + \xi^\mu, \tag{7}$$

which induces a redefinition of the gravitational field tensor

$$\tilde{h}'_{\mu\nu} \equiv \tilde{h}_{\mu\nu} - \partial_\mu\xi_\nu - \partial_\nu\xi_\mu + n_{\mu\nu}\partial_\lambda\xi^\lambda. \tag{8}$$

It can be easily proved that ξ^μ must satisfy the condition

$$\partial_\mu\partial^\mu\xi^\alpha = 0, \tag{9}$$

so that the new gravitational field is in agreement with the Hilbert gauge condition. The solution of Eq. (9) defines the four components of ξ^μ so that the new tensor $\tilde{h}'_{\mu\nu}$ is also a solution of the wave equation (6), and thus has the same physical meaning as $\tilde{h}_{\mu\nu}$. In general, gauge transformations correspond to symmetries of the field equations, which means that the field equations are invariant under such transformations. This implies that the field equations do not determine the field uniquely; however, this ambiguity in determining the field is devoid of any physical meaning.

The simplest solution to the wave equation (6) is a plane wave solution of the form

$$\tilde{h}^{\mu\nu} = A^{\mu\nu} e^{ik_a x^a}, \tag{10}$$

where $A^{\mu\nu}$ is a constant symmetric tensor, the *polarization tensor*, in which information about the amplitude and the polarization of the waves is encoded, while k_α is a constant vector, the *wave vector*, which determines the propagation direction of the wave and its frequency. In physical applications we will use only the real part of the above wave solution. By applying the Hilbert gauge condition on the plane wave solution we obtain the relation $A^{\mu\nu}k_\mu = 0$, the geometrical meaning of which is that $A^{\mu\nu}$ and k_μ are orthogonal. This relation can be written as four equations that impose four conditions on $A^{\mu\nu}$, and this is the first step in reducing the number of its independent components. As a consequence, $A^{\mu\nu}$, instead of having 10 independent components (as has every symmetric second-rank tensor in a four-dimensional space), has only 6 independent ones. Further substitution of the plane wave solution in the wave equation leads to the important equation $k^\mu k_\mu = k_0^2 - (k_x^2 + k_y^2 + k_z^2) = 0$, which means that k_μ is a *lightlike* or *null vector*, i.e., the wave propagates on the light-cone. This means that the speed of the wave is 1, i.e., equal to the speed of light. The frequency of the wave is $\omega = k^0$.

Up to this point it has been proven that $A^{\mu\nu}$ has six arbitrary components, but due to the gauge freedom, i.e., the freedom in choosing the four components of the vector ξ^μ, the actual number of its independent components can be reduced to *two*, in a suitably chosen gauge. The *transverse-traceless* or *TT gauge* is an example of such a gauge. In this gauge, only the spatial components of $\tilde{h}_{\mu\nu}$ are nonzero (hence $\tilde{h}_{\mu 0} = 0$), which means that the wave is *transverse* to its own direction of propagation, and, additionally, the sum of the diagonal components is zero ($\tilde{h}^\mu_\mu = h^0_0 + h^1_1 + h^2_2 + h^3_3 \equiv \tilde{h} = h = 0$) (*traceless*). Due to this last property and Eq. (5), in this gauge there is no difference between $h_{\mu\nu}$ (the perturbation of the metric) and $\tilde{h}_{\mu\nu}$ (the gravitational field). It is customary to write the gravitational wave solution in the TT gauge as $h^{\mathrm{TT}}_{\mu\nu}$. That $A^{\mu\nu}$ has only two independent components means that a gravitational wave is completely described by two dimensionless amplitudes, h_+ and $h_\times$, say. If, for example, we assume a wave propagating along the z direction, then the amplitude $A^{\mu\nu}$ can be written as

$$A^{\mu\nu} = h_+ \varepsilon_+^{\mu\nu} + h_\times \varepsilon_\times^{\mu\nu}, \tag{11}$$

where $\varepsilon_+^{\mu\nu}$ and $\varepsilon_\times^{\mu\nu}$ are the so-called *unit polarization tensors* defined by

$$\varepsilon_+^{\mu\nu} \equiv \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \varepsilon_\times^{\mu\nu} \equiv \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{12}$$

As mentioned earlier, the Riemann tensor is a measure of the curvature of spacetime. A gravitational wave propagating in a flat spacetime, generates periodic distortions which can be described in terms of the Riemann tensor. In linearized theory the Riemann tensor takes the following gauge-independent form

$$R_{\kappa\lambda\mu\nu} = \frac{1}{2}(\partial_{\nu\kappa} h_{\lambda\mu} + \partial_{\lambda\mu} h_{\kappa\nu} - \partial_{\kappa\mu} h_{\lambda\nu} - \partial_{\lambda\nu} h_{\kappa\mu}), \tag{13}$$

which is considerably simplified by choosing the TT gauge:

$$R^{\mathrm{TT}}_{j0k0} = -\frac{1}{2}\frac{\partial^2}{\partial t^2} h^{\mathrm{TT}}_{jk}, \qquad j,k = 1,2,3. \tag{14}$$

Furthermore, in the Newtonian limit

$$R^{\mathrm{TT}}_{j0k0} \approx \frac{\partial^2 \Phi}{\partial x^j \partial x^k}, \tag{15}$$

where Φ describes the gravitational potential in Newtonian theory. Earlier we defined the Riemann tensor as a geometrical object, but this tensor has a simple physical interpretation: it is the tidal force field and describes the relative acceleration between two particles in free fall. If we assume two particles moving freely along geodesics of a curved spacetime with coordinates $x^\mu(\tau)$ and $x^\mu(\tau) + \xi^\mu(\tau)$ [for a given value of the proper time τ, $\xi^\mu(\tau)$ is the displacement vector connecting the two events], it can be shown that, in the case of slowly moving particles,

$$\frac{d^2 \xi^k}{dt^2} \approx -R^{\mathrm{TT}k}_{0j0} \xi^j. \tag{16}$$

This is a simplified form of the equation of *geodesic deviation*. Hence, the tidal force acting on a particle is

$$f^k \approx -m R^k_{0j0} \xi^j, \tag{17}$$

where m is the mass of the particle. Equation (17) corresponds to the standard Newtonian relation for the tidal force acting on a particle in a field Φ.

Keeping this in mind, we will try to visualize the effect of a gravitational wave. Let us first consider two freely falling particles hit by a gravitational wave traveling along the z direction with the (+) polarization present only, i.e.,

$$h^{\mu\nu} = h_+ \varepsilon_+^{\mu\nu} \cos[\omega(t-z)]. \tag{18}$$

Then, the measured distance ξ^x between the two particles originally at a distance ξ_0^x along the x direction, will be

$$\xi^x = \left\{1 - \frac{1}{2} h_+ \cos[\omega(t-z)]\right\} \xi_0^x$$

or

$$\delta\xi^x = \xi^x - \xi_0^x = -\frac{1}{2} h_+ \cos[\omega(t-z)]\xi_0^x, \tag{19}$$

which implies that the relative distance $\delta\xi^x$ between the two particles will oscillate with frequency ω. This does not mean that the particle's coordinate positions change; they remain at rest relative to the coordinates, but the coordinate distance oscillates. If the particles were placed originally along the y direction, the coordinate distance would oscillate according to

$$\xi^y = \left[1 + \frac{1}{2} h_+ \cos[\omega(t-z)]\right] \xi_0^y$$

or

$$\delta\xi^y = \xi^y - \xi_0^y = \frac{1}{2} h_+ \cos[\omega(t-z)]\xi_0^y \tag{20}$$

In other words, the coordinate distances along the two axes oscillate out of phase, that is, when the distance between two particles along the x direction is maximum, the distance of two other particles along the y direction is minimum, and after half a period, it is the other way around. The effects are similar for the other polarization, where the axes along which the oscillations are out of phase are at an angle of 45° with respect to the first ones. This is visualized in Fig. 1, where the effect of a passing gravitational wave on a ring of particles is shown as a series of snapshots closely separated in time.

Another way of understanding the effects of gravitational waves is to study the tidal force field lines. In the TT gauge the equation of the geodesic deviation (16) takes the simple form

$$\frac{d^2\xi^k}{dt^2} \approx \frac{1}{2} \frac{d^2 h_{jk}^{\mathrm{TT}}}{dt^2} \xi^j \tag{21}$$

and the corresponding tidal force is

$$f^k \approx \frac{m}{2} \frac{d^2 h_{jk}^{\mathrm{TT}}}{dt^2} \xi^j. \tag{22}$$

For the wave given by Eq. (18) the two nonzero components of the tidal force are

$$\begin{aligned} f^x &\approx \frac{m}{2} h_+ \omega^2 \cos[\omega(t-z)]\xi_0^x, \\ f^y &\approx -\frac{m}{2} h_+ \omega^2 \cos[\omega(t-z)]\xi_0^y. \end{aligned} \tag{23}$$

It can be easily proved that the divergence of the tidal force is zero ($\partial f^x/\partial\xi^x + \partial f^y/\partial\xi^y = 0$). It can therefore be represented graphically by field lines as in Fig. 2.

Let us now return to the two polarization states represented by the two matrices $\varepsilon_+^{\mu\nu}$ and $\varepsilon_\times^{\mu\nu}$. It is impossible to construct the (+) pattern from the (×) pattern and vice versa; they are orthogonal polarization states. By analogy with electromagnetic waves, the two polarizations could be added with phase difference ($\pm\pi/2$) to obtain circularly polarized waves. The effect of circularly polarized waves on a ring of particles is to deform the ring into a rotating ellipse with either positive or negative helicity. The

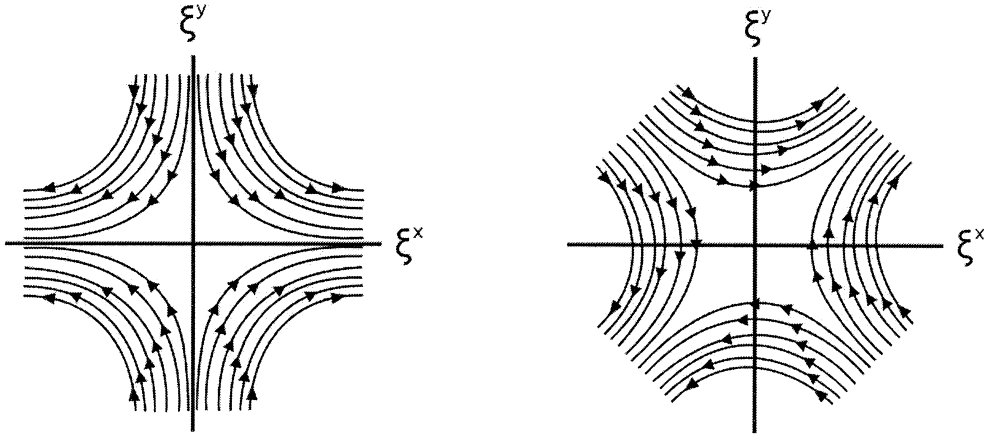


FIGURE 2 The tidal field lines of force for a gravitational wave with polarization (+) (left panel) and (×) (right panel). The orientation of the field lines changes every half period producing the deformations as seen in Fig. 1. Any point accelerates in the direction of the arrows, and the denser are the lines, the stronger is the acceleration. Since the acceleration is proportional to the distance from the center of mass, the force lines get denser as one moves away from the origin. For the polarization (×) the force lines undergo a 45° rotation.

particles themselves do not rotate; they only oscillate in and out around their initial positions. These circularly polarized waves carry angular momentum the amount of which is $(2/\omega)$ times the energy carried by the wave. If we consider the gravitational field, then, according to quantum field theory, the waves are associated with fundamental particles responsible for the gravitational interaction, and a quantum of the field will have energy $\hbar\omega$ and consequently spin $2\hbar$. This means that the quanta of the gravitational field, the gravitons, are spin-2, massless particles (since they travel with the speed of light). Another way of explaining why a graviton should be a spin-2 particle comes from observing Figure 1. One can see that a gravitational wave is invariant under rotations of 180° about its direction of propagation; the pattern repeats itself after half a period. For comparison, electromagnetic waves are invariant under rotations of 360°. In the quantum mechanical description of massless particles, the wavefunction of a particle is invariant under rotations of $360^\circ/s$, where s is the spin of the particle. Thus the photon is a spin-1 particle and the graviton is a spin-2 particle. In other relativistic theories of gravity, the wave field has other symmetries and therefore these theories attribute different spins to the gravitons.

B. Properties of Gravitational Waves

Gravitational waves, once they are generated, propagate almost unimpeded. Indeed, it has been proven that they are even harder to stop than neutrinos! The only significant change they suffer as they propagate is a decrease in amplitude as they travel away from their source and a *redshift* they undergo (cosmological, gravitational, or Doppler), as is the case for electromagnetic waves.

There are other effects that marginally influence the gravitational waveforms, for instance, *absorption* by interstellar or intergalactic matter intervening between the observer and the source, which is extremely weak (actually, the extremely weak coupling of gravitational waves with matter is the main reason that gravitational waves have not been observed). *Scattering* and *dispersion* of gravitational waves are also practically unimportant, although they may have been important during the early phases of the universe (this is also true for absorption). Gravitational waves can be *focused* by strong gravitational fields and also can be *diffracted*, exactly as happens with electromagnetic waves.

There are also a number of "exotic" effects that gravitational waves can experience due to the nonlinear nature of Einstein's equations (purely general-relativistic effects), such as *scattering* by the background curvature, the existence of *tails* of the waves that interact with the waves themselves, *parametric amplification* by the background curvature, *nonlinear coupling of waves with themselves* (creation of *geons*, that is, bundles of gravitational waves held together by their own self-generated curvature), and even *formation of singularities* by colliding waves (for such exotic phenomena see the extensive review by Thorne, 1987). These aspects of nonlinearity affect the majority of gravitational wave sources and from this point of view our understanding of gravitational wave generation is based on approximations. However, the error in these approximations for most processes that generate gravitational waves is expected to be quite small. Powerful numerical codes, using state-of-the-art computer software and hardware, have been developed (and continue to be developed) for minimizing all possible sources of error in order to have as accurate as possible an understanding of the processes that generate gravitational waves and the waveforms produced.

For most of the properties mentioned above there is a correspondence with electromagnetic waves. Gravitational waves are fundamentally different, however, even though they share similar wave properties away from the source. Gravitational waves are emitted by coherent bulk motions of matter (for example, by the implosion of the core of a star during a supernova explosion) or by coherent oscillations of spacetime curvature, and thus they serve as a probe of such phenomena. By contrast, cosmic electromagnetic waves are mainly the result of incoherent radiation by individual atoms or charged particles. As a consequence, from the cosmic electromagnetic radiation we mainly learn about the form of matter in various regions of the universe, especially about its temperature and density, or about the existence of magnetic fields. Strong gravitational waves, are emitted from regions of spacetime where gravity is very strong and the velocities of the bulk motions of matter are near the speed of light. Since most of the time these areas are either surrounded by thick layers of matter that absorb electromagnetic radiation or do not emit any electromagnetic radiation at all (black holes), the only way to study these regions of the universe is via gravitational waves.

C. Energy Flux Carried by Gravitational Waves

Gravitational waves carry energy and cause a deformation of spacetime. The stress-energy carried by gravitational waves cannot be localized within a wavelength. Instead, one can say that a certain amount of stress-energy is contained in a region of space which extends over several wavelengths. It can be proven that in the TT gauge of linearized theory the stress-energy tensor of a gravitational wave (in analogy with the stress-energy tensor of a perfect fluid as defined earlier) is given by

$$t_{\mu\nu}^{\mathrm{GW}} = \frac{1}{32\pi}\left\langle \left(\partial_\mu h_{ij}^{\mathrm{TT}}\right)\left(\partial_\nu h_{ij}^{\mathrm{TT}}\right)\right\rangle, \tag{24}$$

where the angular brackets are used to indicate averaging over several wavelengths. For the special case of a plane wave propagating in the z direction which we considered earlier, the stress-energy tensor has only three nonzero components, which take the simple form

$$t_{00}^{\mathrm{GW}} = \frac{t_{zz}^{\mathrm{GW}}}{c^2} = -\frac{t_{0z}^{\mathrm{GW}}}{c} = \frac{1}{32\pi}\frac{c^2}{G}\omega^2\left(h_+^2 + h_\times^2\right), \tag{25}$$

where t_{00}^{GW} is the energy density, t_{zz}^{GW} is the momentum flux, and t_{0z}^{GW} is the energy flow along the z direction per unit area and unit time (for practical reasons we have restored the normal units). The energy flux has all the properties one would anticipate by analogy with electromagnetic waves: (a) it is conserved (the amplitude dies out as $1/r$, the flux as $1/r^2$), (b) it can be absorbed by detectors, and (c) it can generate curvature like any other energy source in Einstein's formulation of relativity. As an example, by using the above relation, we will estimate the energy flux in gravitational waves from the collapse of the core of a supernova to create a 10-solar-mass black hole at a distance of 50-million light-years (15 Mpc) from the earth (at the distance of the Virgo cluster of galaxies). A conservative estimate of the amplitude of the waves on earth (as we will show later) is of the order of 10^{-22} (at a frequency of about 1 kHz). This corresponds to a flux of about 3 ergs/cm^2 sec. This is an enormous amount of energy flux and is about 10 orders of magnitude larger than the observed energy flux in electromagnetic waves! The basic difference is the duration of the two signals; a gravitational wave signal will last a few milliseconds, whereas an electromagnetic signal lasts many days. This example provides a useful numerical formula for the energy flux:

$$F = 3\left(\frac{f}{1\ \mathrm{kHz}}\right)^2\left(\frac{h}{10^{-22}}\right)^2\frac{\mathrm{ergs}}{\mathrm{cm}^2\ \mathrm{sec}}, \tag{26}$$

from which one can easily estimate the flux on earth, given the amplitude (on earth) and the frequency of the waves.

D. Generation of Gravitational Waves

As mentioned earlier, when the gravitational field is strong there are a number of nonlinear effects on the generation and propagation of gravitational waves. For example, nonlinear effects are significant during the last phases of black hole formation. The analytic description of such a dynamically changing spacetime is impossible, and until numerical relativity provides us with accurate estimates of the dynamics of gravitational fields under such extreme conditions we have to be content with order-of-magnitude estimates. Furthermore, there are differences in the predictions of various relativistic theories of gravity in the case of high concentrations of rapidly varying energy distributions. However, all metric theories of gravity, as long as they admit the correct Newtonian limit, make similar predictions for the total amount of gravitational radiation emitted by "weak" gravitational wave sources, that is, sources where the energy content is small enough to produce only small deformations of the flat spacetime and where all motions are slow compared to the velocity of light.

Let us now try to understand the nature of gravitational radiation by starting from the production of electromagnetic radiation. Electromagnetic radiation emitted by slowly varying charge distributions can be decomposed into a series of multipoles, where the amplitude of the 2^ℓ-pole ($\ell = 0, 1, 2, \ldots$) contains a small factor a^ℓ, with a equal to the ratio of the diameter of the source to the typical wavelength, namely, a number typically much smaller than 1. From this point of view the strongest electromagnetic radiation would be expected for monopolar radiation ($\ell = 0$), but this is completely absent because the electromagnetic monopole moment is proportional to the total charge, which does not change with time (it is a conserved quantity). Therefore, electromagnetic radiation consists only of $\ell \geq 1$ multipoles, the strongest being the electric dipole radiation ($\ell = 1$), followed by the weaker magnetic dipole and electric quadrupole radiation ($\ell = 2$). One could proceed with a similar analysis for gravitational waves and by following the same arguments show that mass conservation (which is equivalent to charge conservation in electromagnetic theory) will exclude monopole radiation. Also, the rate of change of the mass dipole moment is proportional to the linear momentum of the system, which is a conserved quantity, and therefore there cannot be any mass dipole radiation in Einstein's relativity theory. The next strongest form of electromagnetic radiation is the magnetic dipole. For the case of gravity, the change of the "magnetic dipole" is proportional to the angular momentum of the system, which is also a conserved quantity, and thus there is no dipolar gravitational radiation of *any* sort. It follows that gravitational radiation is of quadrupolar or higher nature and is directly linked to the quadrupole moment of the mass distribution.

As early as 1918, Einstein derived the quadrupole formula for gravitational radiation. This formula states that the wave amplitude h_{ij} is proportional to the second time derivative of the quadrupole moment of the source:

$$h_{ij} = \frac{2}{r}\frac{G}{c^4}\ddot{Q}_{ij}^{\mathrm{TT}}\left(t - \frac{r}{c}\right), \tag{27}$$

where

$$Q_{ij}^{\mathrm{TT}}(x) = \int \rho\left(x^i x^j - \frac{1}{3}\delta^{ij} r^2\right) d^3x \qquad (28)$$

is the quadrupole moment in the TT gauge evaluated at the retarded time $t - r/c$ and ρ is the matter density in a volume element d^3x at the position x^i. This result is quite accurate for all sources as long as the reduced wavelength $\lambda\!\!\!^{-} = \lambda/2\pi$ is much longer than the source size R. It should be pointed out that the above result can be derived via a quite cumbersome calculation in which we solve the wave equation (6) with a source term $T_{\mu\nu}$ on the right-hand side. In the course of such a derivation, a number of assumptions must be used. In particular, the observer must be located at a distance $r \gg \lambda\!\!\!^{-}$, far greater than the reduced wavelength (in what is called the radiation zone) and $T_{\mu\nu}$ must not change very quickly.

Using the formulas (24) and (25) for the energy carried by gravitational waves, one can derive the luminosity in gravitational waves as a function of the third-order time derivative of the quadrupole moment tensor. This is the quadrupole formula

$$L_{\mathrm{GW}} = -\frac{dE}{dt} = \frac{1}{5}\frac{G}{c^5}\langle \dddot{Q}_{ij} \dddot{Q}_{ij} \rangle. \qquad (29)$$

Based on this formula, we derive some additional formulas which provide order-of-magnitude estimates for the amplitude of the gravitational waves and the corresponding power output of a source. First, the quadrupole moment of a system is approximately equal to the mass M of the part of the system that moves, times the square of the size R of the system. This means that the third-order time derivative of the quadrupole moment is

$$\dddot{Q}_{ij} \sim MR^2/T^3 \sim M\upsilon^2/T \sim E_{\mathrm{ns}}/T, \qquad (30)$$

where υ is the mean velocity of the moving parts, E_{ns} is the kinetic energy of the component of the source's internal motion that is nonspherical, and T is the time scale for a mass to move from one side of the system to the other. The time scale (or period) is actually proportional to the inverse of the square root of the mean density of the system

$$T \sim \sqrt{R^3/GM}. \qquad (31)$$

This relation provides a rough estimate of the characteristic frequency of the system $f = 2\pi/T$. Then the luminosity of gravitational waves of a given source is approximately

$$L_{\mathrm{GW}} \sim \frac{G}{c^5}\left(\frac{M}{R}\right)^5 \sim \frac{G}{c^5}\left(\frac{M}{R}\right)^2 \upsilon^6 \sim \frac{c^5}{G}\left(\frac{R_{\mathrm{Sch}}}{R}\right)^2 \left(\frac{\upsilon}{c}\right)^6, \qquad (32)$$

where $R_{\mathrm{Sch}} = 2GM/c^2$ is the Schwarzschild radius of the source. It is obvious that the maximum value of the luminosity in gravitational waves can be achieved if the source's dimensions are of the order of its Schwarzschild radius and the typical velocities of the components of the system are of the order of the speed of light. This explains why we expect the best gravitational wave sources to be highly relativistic compact objects. The above formula sets also an upper limit on the power emitted by a source, which for $R \sim R_{\mathrm{Sch}}$ and $\upsilon \sim c$ is

$$L_{\mathrm{GW}} \sim c^5/G = 3.6 \times 10^{59}\ \mathrm{ergs/sec}. \qquad (33)$$

This is an immense power, often called the *luminosity of the universe*.

Using the above order-of-magnitude estimates, we can get a rough estimate of the amplitude of gravitational waves at a distance r from the source:

$$h \sim \frac{G}{c^4}\frac{E_{\mathrm{ns}}}{r} \sim \frac{G}{c^4}\frac{\varepsilon E_{\mathrm{kin}}}{r}, \qquad (34)$$

where $\varepsilon E_{\mathrm{kin}}$ (with $0 \le \varepsilon \le 1$) is the fraction of kinetic energy of the source that is able to produce gravitational waves. The factor ε is a measure of the asymmetry of the source and implies that only a time-varying quadrupole moment will emit gravitational waves. For example, even if a huge amount of kinetic energy is involved in a given explosion and/or implosion, if the event takes place in a spherically symmetric manner, there will be no gravitational radiation.

Another formula for the amplitude of gravitational waves can be derived from the flux formula (26). If, for example, we consider an event (perhaps a supernova explosion) at the Virgo cluster during which the energy equivalent of 10^{-4} solar masses is released in gravitational waves at a frequency of 1 kHz and with signal duration of the order of 1 msec, the amplitude of the gravitational waves on earth will be

$$h \approx 10^{-22}\left(\frac{E}{10^{-4}M_{\mathrm{Sun}}}\right)^{1/2}\left(\frac{f}{1\ \mathrm{kHz}}\right)^{-1} \times \left(\frac{\tau}{1\ \mathrm{msec}}\right)^{-1/2}\left(\frac{r}{15\ \mathrm{Mpc}}\right)^{-1}. \qquad (35)$$

These numbers explain why experimenters are trying so hard to build ultrasensitive detectors.

Finally, it is useful to know the damping time, that is, the time it takes for a source to transform a fraction $1/e$ of its energy into gravitational radiation. One can obtain a rough estimate from the formula

$$\tau = \frac{E_{\mathrm{kin}}}{L_{\mathrm{GW}}} \sim \frac{1}{c} R\left(\frac{R}{R_{\mathrm{Sch}}}\right)^3. \qquad (36)$$

For example, for a non-radially oscillating neutron star with a mass of roughly 1.4 solar masses and a radius of 12 km, the damping time will be of the order of 1 msec. Also, by using formula (31), we get an estimate for the

frequency of oscillation which is directly related to the frequency of the emitted gravitational waves, roughly 2 kHz for the above case.

E. Rotating Binary System

Among of the most interesting sources of gravitational waves are binaries. The inspiraling of such systems, consisting of black holes or neutron stars, is, as we will discuss later, the most promising source for gravitational wave detectors. Binary systems are also sources of the gravitational waves whose dynamics we understand the best. If we assume that the two bodies making up the binary lie in the x–y plane and their orbits are circular (see Fig. 3), then the only nonvanishing components of the quadrupole tensor are

$$Q_{xx} = -Q_{yy} = \frac{1}{2}\mu a^2 \cos 2\Omega t,$$
$$Q_{xy} = Q_{yx} = \frac{1}{2}\mu a^2 \sin 2\Omega t, \tag{37}$$

where Ω is the orbital angular velocity, $\mu = M_1 M_2/M$ is the reduced mass of the system, and $M = M_1 + M_2$ is its total mass.

According to Eq. (29) the gravitational radiation luminosity of the system is

$$L^{\mathrm{GW}} = \frac{32}{5}\frac{G}{c^5}\mu^2 a^4 \Omega^6 = \frac{32}{5}\frac{G^4}{c^5}\frac{M^3\mu^2}{a^5}, \tag{38}$$

where, in order to obtain the last part of the relation, we have used Kepler's third law, $\Omega^2 = GM/a^3$. As the gravitating system loses energy by emitting radiation, the distance between the two bodies shrinks at a rate

$$\frac{da}{dt} = -\frac{64}{5}\frac{G^3}{c^5}\frac{\mu M^2}{a^3}, \tag{39}$$

and the orbital frequency increases accordingly ($\dot{T}/T = 1.5\dot{a}/a$). If, for example, the present separation of the two stars is a_0, then the binary system will coalesce after a time

$$\tau = \frac{5}{256}\frac{c^5}{G^3}\frac{a_0^4}{\mu M^2}. \tag{40}$$

Finally, the amplitude of the gravitational waves is

$$h = 5 \times 10^{-22}\left(\frac{M}{2.8M_{\mathrm{Sun}}}\right)^{2/3}\left(\frac{\mu}{0.7M_{\mathrm{Sun}}}\right)$$
$$\times\left(\frac{f}{100\ \mathrm{Hz}}\right)^{2/3}\left(\frac{15\ \mathrm{Mpc}}{r}\right). \tag{41}$$

In all these formulas we have assumed that the orbits are circular. In general, the orbits of the two bodies are approximately ellipses, but it has been shown that long before the coalescence of the two bodies, the orbits become circular, at least for long-lived binaries, due to gravitational radiation. Also, the amplitude of the emitted gravitational waves depends on the angle between the line of sight and the axis of angular momentum; formula (41) refers to an observer along the axis of the orbital angular momentum. The complete formula for the amplitude contains angular factors of order 1. The relative strength of the two polarizations depends on that angle as well.

If three or more detectors observe the same signal, it is possible to reconstruct the full waveform and deduce many details of the orbit of the binary system.

As an example, we will provide some details of the well-studied pulsar PSR 1913 + 16 (the Hulse–Taylor pulsar), which is expected to coalesce after $\sim 3.5 \times 10^8$ years. The binary system is roughly 5 kpc away from earth, the masses of the two neutron stars are estimated to be $\sim$1.4 solar masses each, and the present period of the system is $\sim$7 hr, 45 min. The predicted rate of period change is

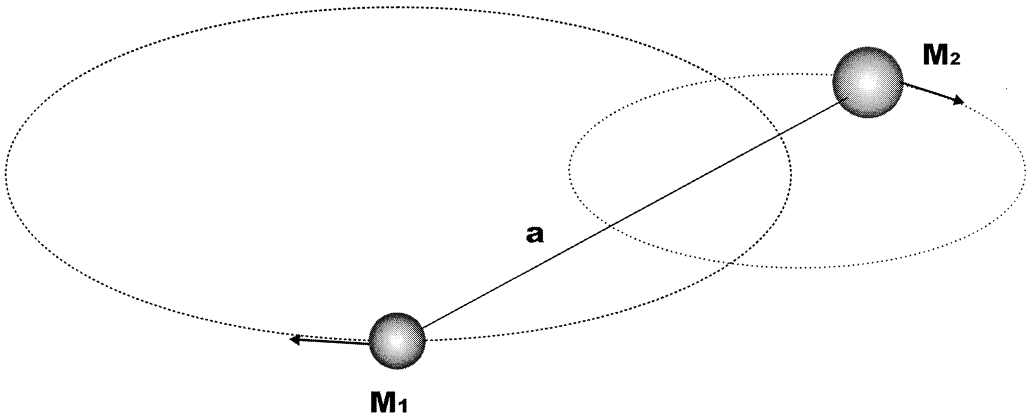


FIGURE 3 A system of two bodies orbiting around their common center of gravity. Binary systems are the best sources of gravitational waves.

$\dot{T} = -2.40 \times 10^{-12}$ sec/sec, while the corresponding observed value is in excellent agreement with the predictions, i.e., $\dot{T} = (-2.30 \pm 0.22) \times 10^{-12}$ sec/sec; finally the present amplitude of gravitational waves is of the order of $h \sim 10^{-23}$ at a frequency of $\sim 7 \times 10^{-5}$ Hz.

III. DETECTION OF GRAVITATIONAL RADIATION

The first attempt to detect gravitational waves was undertaken by Joseph Weber during the early 1960s. He developed the first resonant mass detector and inspired many other physicists to build new detectors and explore from a theoretical viewpoint possible cosmic sources of gravitational radiation.

A pair of masses joined by a spring can be viewed as the simplest conceivable detector; see Figure 4. In practice, a cylindrical massive metal bar or even a massive sphere is used instead of this simple system. When a gravitational wave hits such a device, it causes the bar to vibrate. By monitoring this vibration, we can reconstruct the true waveform. The next step following resonant mass detectors was the replacement of the spring by pendulums. In this new detector the motions induced by a passing gravitational wave would be detected by monitoring, via laser interferometry, the relative change in the distance of two freely suspended bodies. The use of interferometry is probably the most decisive step in our attempt to detect gravitational wave signals. In what follows, we will discuss both resonant bars and laser interferometric detectors.

Although the basic principle of such detectors is very simple, the sensitivity of detectors is limited by various sources of noise. The internal noise of the detectors can be Gaussian or non-Gaussian. The non-Gaussian noise may occur several times per day such as by strain releases in the suspension systems which isolate the detector from any environmental mechanical source of noise, and the only way to remove this type of noise is via comparisons of the data streams from various detectors. The so-called Gaussian noise obeys the probability distribution of Gaussian statistics and can be characterized by a *spectral density* $S_n(f)$. The observed signal at the output of a detector consists of the true gravitational wave strain h and Gaussian noise. The optimal method to detect a gravitational wave signal leads to the following signal-to-noise ratio:

$$\left(\frac{S}{N}\right)^2_{\text{opt}} = 2\int_0^\infty \frac{|\tilde{h}(f)|^2}{S_n(f)}\,df, \tag{42}$$

where $\tilde{h}(f)$ is the Fourier transform of the signal waveform. It is clear from this expression that the sensitivity of gravitational wave detectors is limited by noise.



FIGURE 4 A pair of masses joined by a spring can be viewed as the simplest conceivable detector.

A. Resonant Detectors

Suppose that a gravitational wave propagating along the z axis with pure (+) polarization impinges on an idealized detector, two masses joined by a spring along the x axis as in Fig. 4. We will assume that $h_{\mu\nu}$ describes the strain produced in the spacetime by the passing wave. We will try to calculate the amplitude of the oscillations induced on the spring detector by the wave and the amount of energy absorbed by this detector. The tidal force induced on the detector is given by Eq. (23), and the masses will move according to the following equation of motion:

$$\ddot{\xi} + \dot{\xi}/\tau + \omega_0^2\xi = -\tfrac{1}{2}\omega^2 L h_+ e^{i\omega t}, \tag{43}$$

where ω_0 is the natural vibration frequency of the detector, τ is the damping time of the oscillator due to frictional forces, L is the separation between the two masses, and ξ is the relative change in the distance of the two masses. The gravitational wave plays the role of the driving force for the ideal oscillator, and the solution to the above equation is

$$\xi = \frac{\frac{1}{2}\omega^2 L h_+ e^{i\omega t}}{\omega_0^2 - \omega^2 + i\omega/\tau}. \tag{44}$$

If the frequency ω of the impinging wave is near the natural frequency ω_0 of the oscillator (near resonance), the detector is excited into large-amplitude motions and it rings like a bell. Actually, in the case of $\omega = \omega_0$, we get the maximum amplitude $\xi_{\max} = \omega_0 \tau L h_+/2$. Since the size of the detector L and the amplitude of the gravitational waves h_+ are fixed, large-amplitude motions can be achieved only by increasing the quality factor $Q = \omega_0\tau$ of the detector. In practice, the frequency of the detector is fixed by its size and the only improvement we can get is by choosing the type of material so that long relaxation times are achieved.

The *cross section* is a measure of the interception ability of a detector. For the special case of resonance, the average cross section of the test detector, assuming any possible direction of the wave, is

$$\sigma = \frac{32\pi}{15}\frac{G}{c^3}\omega_0 Q M L^2. \tag{45}$$

This formula is general; it applies even if we replace our toy detector with a massive metal cylinder, for example, Weber's first detector. That detector had the following characteristics: mass M = 1410 kg, length L = 1.5 m, diameter 66 cm, resonant frequency $\omega_0 = 1660\,\mathrm{Hz}$, and quality factor $Q = \omega_0\tau = 2 \times 10^5$. For these values the calculated cross section is roughly $3 \times 10^{-19}\,\mathrm{cm}^2$, which is quite small, and even worse, it can be reached only when the frequency of the impinging wave is very close to the resonance frequency (the typical resonance width is usually of the order of 0.1–1 Hz).

In reality, the efficiency of a resonance bar detector depends on several other parameters. Here, we will discuss only the more fundamental ones. Assuming perfect isolation of the resonant bar detector from any external source of noise (acoustical, seismic, electromagnetic), the thermal noise is the only factor limiting our ability to detect gravitational waves. Thus, in order to detect a signal, the energy deposited by the gravitational wave every τ seconds should be larger than the energy kT due to thermal fluctuations. This leads to a formula for the minimum detectable energy flux of gravitational waves, which, following Eq. (25), leads into a minimum detectable strain amplitude h,

$$h \geq \frac{1}{\omega_0 L Q}\sqrt{\frac{15kT}{M}}. \tag{46}$$

For Weber's detector, at room temperature this yields a minimum detectable strain of the order of 10^{-20}. However, this estimate of the minimum sensitivity applies only to gravitational waves whose duration is at least as long as the damping time of the bar's vibrations and whose frequency perfectly matches the resonant frequency of the detector. For burst signals or for periodic signals which are off-resonance (with regard to the detector) the sensitivity of a resonant bar detector decreases further by several orders of magnitude.

In reality, modern resonant bar detectors are quite complicated devices, consisting of a solid metallic cylinder weighing a few tons and suspended *in vacuo* by a cable wrapped under its center of gravity (Figure 5). This suspension system protects the antenna from external mechanical shocks. The whole system is cooled down to temperatures of a few kelvins or even millikelvins. To monitor the vibrations of the bar, piezoelectric transducers (or the more modern capacitive ones) are attached to the bar. The transducers convert the bar's mechanical energy into electrical energy. The signal is amplified by an ultralow-frequency amplifier by using a device called a SQUID (superconducting quantum interference device) before it becomes available for data analysis. Transducers and amplifiers of electronic signals require careful design to achieve low noise combined with adequate signal transfer.

The above description of the resonant bar detectors shows that, in order to achieve high sensitivity, one has to:

1. *Create more massive antennas.* Today, most antennas are about 50% more massive than Weber's early antenna. There are studies and research plans for future construction of spherical antennas weighing up to 100 tons.
2. *Obtain higher quality factor Q.* Modern antennas generally use aluminum alloy 5056 ($Q \sim 4 \times 10^7$); niobium (which is used in the Niobe detector) is even better ($Q \sim 10^8$), but much more expensive. Silicon

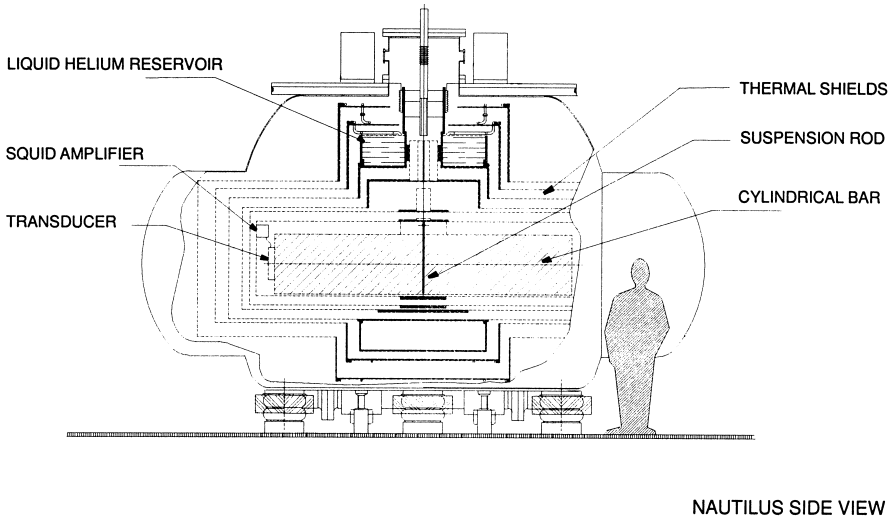


FIGURE 5 A sketch of Nautilus at Frascati, near Rome, probably the most sensitive resonant detector available.

or sapphire bars would enhance the quality factor even more, but experimenters must first find a way to produce large, single pieces of these crystals.

3. *Lower the temperature of the antenna as much as possible*. Advanced cryogenic techniques have been used and the resonant bar detectors are probably the coldest places in the universe. Typical cooling temperatures for the most advanced antennas are below the temperature of liquid helium.
4. *Achieve strong coupling between the antenna and the electronics and low electrical noise*. The bar detectors include the best available technology related to transducers and integrate the most recent advances in SQUID technology.

Since the early 1990s, a number of resonant bar detectors have been in nearly continuous operation in several places around the world. They have achieved sensitivities of a few times 10^{-21}, but there has been no clear evidence of gravitational wave detection. As we will discuss later, they will have a good chance of detecting a gravitational wave signal from a supernova explosion in our galaxy, although, this is a rather rare event (2–5 per century).

The details of the most sensitive cryogenic bar detectors in operation are as follows:

- Allegro (Baton Rouge, LA): Mass 2296 kg (Aluminum 5056), length 3 m, bar temperature 4.2 K, mode frequency 896 Hz.
- Auriga (Legrano, Italy): Mass 2230 kg (Aluminum 5056), length 2.9 m, bar temperature 0.2 K, mode frequency 913 Hz.
- Explorer (CERN, Switzerland): Mass 2270 kg (Aluminum 5056), length 3 m, bar temperature 2.6 K, mode frequency 906 Hz.
- Nautilus (Frascati, Italy): Mass 2260 kg (Aluminum 5056), length 3 m, bar temperature 0.1 K, mode frequency 908 Hz.
- Niobe (Perth, Australia): Mass 1500 kg (niobium), length 1.5 m, bar temperature 5 K, mode frequency 695 Hz.

Also, there are plans for construction of massive *spherical* resonant detectors, the advantages of which will be their high mass (100 tons), their broader sensitivity (up to 100–200 Hz), and their omnidirectional sensitivity. In a spherical detector, five modes at a time will be excited, which is equivalent to five independent detectors oriented in different ways. This offers the opportunity, in the case of detection, to obtain direct information about the polarization of the wave and the direction to the source.

B. Beam Detectors

1. Laser Interferometers

A laser interferometer is an alternative gravitational wave detector that offers the possibility of very high sensitivities over a broad frequency band. Originally, the idea was to construct a new type of resonant detector with much larger dimensions. As one can realize from the relations (45) and (46), the longer the resonant detector is, the more sensitive it becomes. One could then try to measure the relative change in the distance of two well-separated masses by monitoring their separation via a laser beam that continuously bounces back and forth between them. (This technique is actually used in searching for gravitational waves by using the so-called Doppler tracking technique, where a distant interplanetary spacecraft is monitored from earth through a microwave tracking link; the earth and spacecraft act as free particles.) Soon, it was realized that it is much easier to use laser light to measure relative changes in the lengths of two perpendicular arms; see Figure 6. Gravitational waves that are propagating perpendicular to the plane of the interferometer will increase the length of one arm of the interferometer and shorten the length of the other arm. This technique of monitoring waves is based on Michelson interferometry. L-shaped interferometers are particularly suited to the detection of gravitational waves due to their quadrupolar nature.

Figure 6 shows a schematic design of a Michelson interferometer; the three masses M_0, M_1, and M_2 are freely suspended. Note that the resonant frequencies of these pendulums should be much smaller than the frequencies of the waves that we are supposed to detect since the pendulums are supposed to behave like free masses. Mirrors are attached to M_1 and M_2, and the mirror attached to mass M_0 splits the light into two perpendicular directions. The light is reflected at the two corner mirrors and returns to the beam splitter. The splitter now half-transmits and half-reflects each of the beams. One part of each beam goes back to the laser, and the other parts are combined to reach the photodetector, where the fringe pattern is monitored. If a gravitational wave slightly changes the lengths of the two arms, the fringe pattern will change, and so by monitoring the changes of the fringe pattern one can measure the changes in the arm lengths and consequently monitor the incoming gravitational radiation.[1]

Let us consider an impinging gravitational wave with amplitude h and $(+)$ polarization propagating perpendicular to the plane of the detector. We will further assume that

[1]In practice, things are arranged so that all light that returns on the beam splitter from the corner mirrors is sent back into the laser, and only if there is some motion of the masses there is an output at the photodetector.

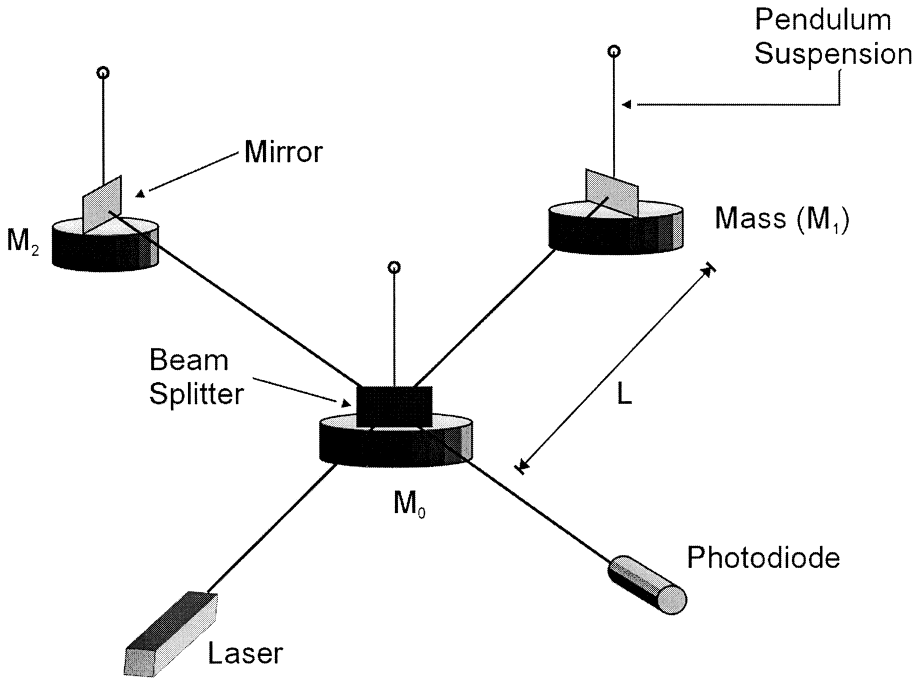


FIGURE 6 Schematic design of a Michelson interferometer.

the frequency is much higher than the resonant frequency of the pendulums and the wavelength is much longer than the arm length of the detector. Such a wave will generate a change of $\Delta L \sim hL/2$ in the arm length along the x direction and an opposite change in the arm length along the y direction, according to equations (19) and (20). The total difference in length between the two arms will be

$$\frac{\Delta L}{L} \sim h. \tag{47}$$

For a gravitational wave with amplitude $h \sim 10^{-21}$ and detector arm length 4 km (such as LIGO), this will induce a change in the arm length of about $\Delta L \sim 10^{-16}$ cm. In the general case, when a gravitational wave with arbitrary polarization impinges on the detector from a random direction, the above formula will be modified by some angular coefficients of order 1.

If the light bounces a few times between the mirrors before it is collected in the photodiode, the effective arm length of the detector is increased considerably and the measured variations of the arm lengths will be increased accordingly. This is a quite efficient procedure for making the arm length longer. For example, a gravitational wave at a frequency of 100 Hz has a wavelength of 3000 km, and if we assume 100 bounces of the laser beam in the arms of the detector, the effective arm length of the detector is 100 times larger than the actual arm length, but still this is 10 times smaller than the wavelength of the incoming wave. The optical cavity that is created between the mirrors of the detector is known as a Fabry–Perot cavity and is used in modern interferometers.

In the remainder of this subsection we will focus on the Gaussian sources of noise and their expected influence on the sensitivity of laser intereferometers.

a. Photon shot noise. When a gravitational wave produces a change ΔL in the arm length, the phase difference between the two light beams changes by an amount $\Delta\varphi = 2b\Delta L/\lambda\!\!\!^{-}$, where $\lambda\!\!\!^{-}$ is the reduced wavelength of the laser light ($\sim 10^{-8}$ cm) and b is the number of bounces of the light in each arm. It is expected that a detectable gravitational wave will produce a phase shift of the order of 10^{-9} rad. The precision of the measurements, though, is ultimately restricted by fluctuations in the fringe pattern due to fluctuations in the number of detected photons. The number of photons that reach the detector is proportional to the intensity of the laser beam and can be estimated via the relation $N = N_0 \sin^2(\Delta\varphi/2)$, where N_0 is the number of photons that the laser supplies and N is the number of detected photons. Inversion of this equation leads to an estimation of the relative change of the arm lengths ΔL by measuring the number of the emerging photons N. However, there are statistical fluctuations in the population of photons, which are proportional to the square root of the number of photons. This implies an uncertainty in the measurement of the arm length

$$\delta(\Delta L) \sim \frac{\lambda\!\!\!^{-}}{2b\sqrt{N_0}}. \tag{48}$$

Thus, the minimum gravitational wave amplitude that we can measure is

$$h_{\min} = \frac{\delta(\Delta L)}{L} = \frac{\Delta L}{L} \sim \frac{\lambda\!\!\!^{-}}{bLN_0^{1/2}} \sim \frac{1}{bL}\left(\frac{\hbar c \lambda\!\!\!^{-}}{\tau I_0}\right)^{1/2}, \tag{49}$$

where I_0 is the intensity of the laser light (∼5–10 W) and τ is the duration of the measurement. This limitation in the detector's sensitivity due to the photon counting uncertainty is known as *photon shot noise*. For a typical laser interferometer the photon shot noise is the dominant source of noise for frequencies above 200 Hz, while its power spectral density $S_n(f)$ for frequencies 100−200 Hz is of the order of $\sim 3 \times 10^{-23}\sqrt{\text{Hz}}$.

b. Radiation pressure noise. According to formula (49), the sensitivity of a detector can be increased by increasing the intensity of the laser. However, a very powerful laser produces a large radiation pressure on the mirrors. Then an uncertainty in the measurement of the momentum deposited on the mirrors leads to a proportional uncertainty in the position of the mirrors or, equivalently, in the measured change in the arm lengths. Then, the minimum detectable strain is limited by

$$h_{\min} \sim \frac{\tau}{m}\frac{b}{L}\left(\frac{\tau\hbar I_0}{c\lambda\!\!\!^{-}}\right)^{1/2}, \tag{50}$$

where m is the mass of the mirrors. As we have seen, the photon shot noise decreases as the laser power increases, while the inverse is true for the noise due to radiation pressure fluctuations. If we try to minimize these two types of noise with respect to the laser power, we get a minimum detectable strain for the optimal power via the very simple relation

$$h_{\min} \sim \frac{1}{L}\left(\frac{\tau\hbar}{m}\right)^{1/2}, \tag{51}$$

which for the LIGO detector (where the mass of the mirrors is ∼100 kg and the arm length is 4 km), for observation time of 1 msec gives $h_{\min} \approx 10^{-23}$.

c. Quantum limit. An additional source of uncertainty in the measurements is set by Heisenberg's principle, which says that the knowledge of the position and the momentum of a body is restricted by the relation $\Delta x \cdot \Delta p \geq \hbar$. For an observation that lasts some time τ, the smallest measurable displacement of a mirror of mass m is ΔL; assuming that the momentum uncertainty is $\Delta p \approx m \cdot \Delta L/\tau$, we get a minimum detectable strain due to quantum uncertainties

$$h_{\min} = \frac{\Delta L}{L} \sim \frac{1}{L}\left(\frac{\tau\hbar}{m}\right)^{1/2}. \tag{52}$$

Surprisingly, this is identical to the optimal limit that we calculated earlier for the other two types of noise. The standard quantum limit does set a fundamental limit on the sensitivity of beam detectors. An interesting feature of the quantum limit is that it depends only on a single parameter, the mass of the mirrors.

d. Seismic noise. At frequencies below 60 Hz, the noise in the interferometers is dominated by seismic noise. This noise is due to geological activity of the earth and human sources, e.g., traffic and explosions. The vibrations of the ground couple to the mirrors via the wire suspensions which support them. This effect is strongly suppressed by properly designed suspension systems. Still, seismic noise is very difficult to eliminate at frequencies below 5–10 Hz.

e. Residual gas-phase noise. The statistical fluctuations of the residual gas density induce a fluctuation of the refraction index and consequently of the monitored phase shift. Hence, the residual gas pressure through which the laser beams travel should be extremely low. For this reason the laser beams are enclosed in pipes over their entire length. Inside the pipes a high vacuum of the order of 10^{-9} torr guarantees elimination of this type of noise.

Prototype laser interferometric detectors have been constructed in the United States, Germany, and the United Kingdom. These detectors have an arm length of a few tens of meters and they have achieved sensitivities of the order of $h \sim 10^{-19}$. A new generation of laser interferometric detectors is under construction and their operation will start by the year 2001, with the first science run to commence around 2002–2003. The American LIGO (Laser Interferometer Gravitational Observatory) project consists of two detectors with arm length of 4 km, one in Hanford, Washington, one in Livingston, Louisiana. The detector in Hanford includes, in the same vacuum system, a second detector with an arm length of 2 km.

The Italian/French Virgo detector of arm length 3 km at Cascina near Pisa, Italy, is designed to have better sensitivity at lower frequencies. GEO600 is a German/British detector built in Hannover, Germany. It has a 600 m arm length and is going to be in operation roughly at the same time as LIGO. The completed TAMA300 detector in Tokyo has an arm length of 300 m and is at an advanced stage of testing of the various components.

2. Space Detectors

Both bar and laser interferometers are high-frequency detectors, but there are a number of interesting gravitational

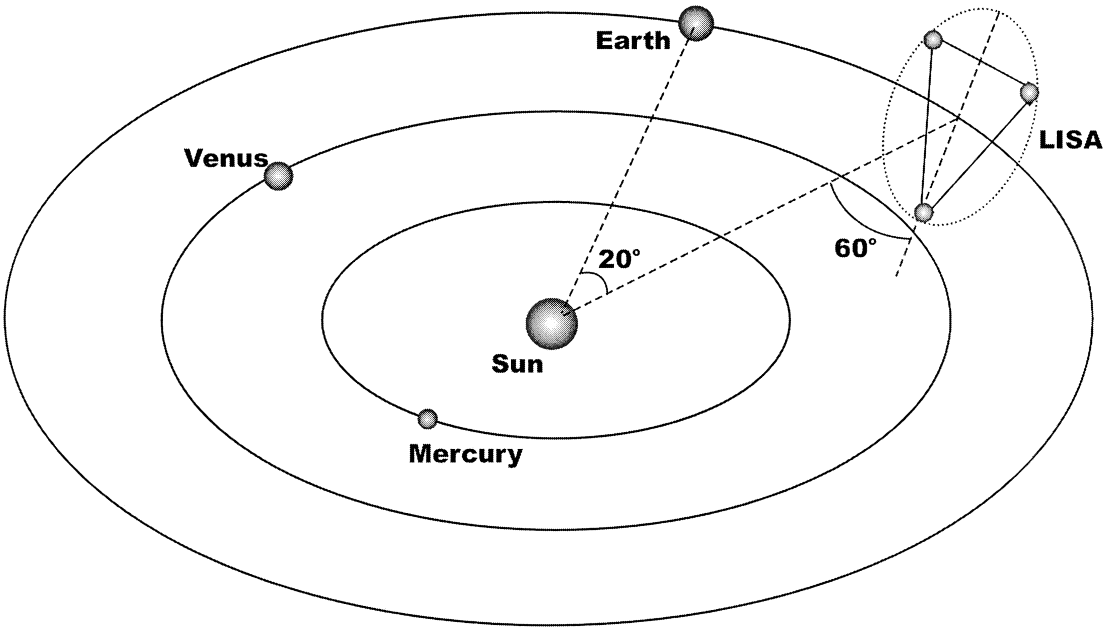


FIGURE 7 Schematic design of the space interferometer LISA.

wave sources which emit signals at lower frequencies. The seismic noise provides an insurmountable obstacle in any earth-based experiment and the only way to overcome this barrier is to fly a laser interferometer in space. LISA (Laser Interferometer Space Antenna) is such a system. It has been proposed by European and American scientists and has been adopted by the European Space Agency (ESA) as a cornerstone mission; recently NASA joined the effort. The launch date is expected to be around 2008.

LISA will consist of three identical drag-free spacecraft forming an equilateral triangle with one spacecraft at each vertex (Fig. 7). The distance between the two vertices (the arm length) is 5×10^6 km. The spacecraft will be placed into the same heliocentric orbit as earth, but about 20° behind earth. The equilateral triangle will be inclined at an angle of 60° with respect to earth's orbital plane. The three spacecraft will track each other optically by using laser beams. Because of the diffraction losses it is not feasible to reflect the beams back and forth as is done with LIGO. Instead, each spacecraft will have its own laser. The lasers will be phase locked to each other, achieving the same kind of phase coherence as LIGO does with mirrors. The configuration will function as three partially independent and partially redundant gravitational wave interferometers.

At frequencies $f \geq 10^{-3}$ Hz, LISA's noise is mainly due to photon shot noise. The sensitivity curve steepens at $f \sim 3 \times 10^{-2}$ Hz because at larger frequencies the gravitational wave's period is shorter than the round-trip light travel time in each arm. For $f \leq 3 \times 10^{-2}$ Hz, the noise is due to buffeting-induced random motions of the spacecraft and cannot be removed by the drag-compensation system. LISA's sensitivity is roughly the same as that of LIGO, but at 10^5 times lower frequency. Since the gravitational wave energy flux scales as $F \sim f^2 h^2$, this corresponds to 10^{10} times better energy sensitivity than LIGO.

3. Satellite Tracking

The Doppler delay of communication signals between earth-based stations and spacecraft underlies another type of gravitational wave detector. A radio signal of frequency v_0 is transmitted to a spacecraft and is coherently transported back to earth, where it is received and its frequency measured with a highly stable clock (typically a hydrogen maser). The relative change $\Delta v/v_0$ as function of time is monitored. A gravitational wave propagating through the solar system causes small perturbations in $\Delta v/v_0$. The relative shift in the frequency of the signals is proportional to the amplitude of gravitational waves. With this technique, broad-band searches are possible in the millihertz frequency band, and thanks to very stable atomic clocks it is possible to achieve sensitivities of order $h_{\min} \sim 10^{-13} - 10^{-15}$. Noise sources that affect the sensitivity of Doppler tracking experiments can be divided into two broad classes: (a) instrumental and (b) related to propagation. At the high-frequency end of the band accessible to Doppler tracking, thermal noise dominates over all other noise sources, typically at about 0.1 Hz. Among all other sources of instrumental noise (transmitter and receiver,

mechanical stability of the antenna, stability of the spacecraft etc), clock noise has been shown to be the most important instrumental source of frequency fluctuations. The propagation noise is due to fluctuations in the index of refraction of the troposphere, ionosphere, and interplanetary solar plasma. Both NASA and ESA have performed such measurements and there is continued effort in this direction.

4. Pulsar Timing

Pulsars are extremely stable clocks and by measuring irregularities in their pulses we expect to set upper limits on background gravitational waves (see next section). If an observer monitors simultaneously two or more pulsars, the correlation of their signals can be used to detect gravitational waves. Since such observation requires time scales of the order of 1 year, this means that the waves have to be of extremely low frequencies.

IV. ASTRONOMICAL SOURCES OF GRAVITATIONAL WAVES

The new generation of gravitational wave detectors (LIGO, Virgo) have very good chances of detecting gravitational waves, but until these expectations are fulfilled, we can only make educated guesses as to the possible astronomical sources of gravitational waves. The detectability of these sources depends on three parameters: their intrinsic gravitational wave luminosity, their event rate, and their distance from the earth. The luminosity can be approximately estimated via the quadrupole formula discussed earlier. Even though there are certain restrictions in its applicability (weak field, slow motion), it provides a very good order-of-magnitude estimate for the expected gravitational wave flux on earth. The rate at which various events with high luminosity in gravitational waves take place is extrapolated from astronomical observations in the electromagnetic spectrum. Still, there might be a number of *gravitationally* luminous sources, for example, binary black holes, for which we have no direct observations in the electromagnetic spectrum. Finally, the amplitude of gravitational wave signals decreases as one over the distance to the source. Thus, a signal from a supernova explosion might be clearly detectable if the event takes place in our galaxy (2–3 events per century) but it is highly unlikely to be detected if the supernova explosion occurs at far greater distances, of order 100 Mpc, where the event rate is high and at least a few events per day take place. All three factors have to be taken into account when discussing sources of gravitational waves, but we will not discuss this matter further, as this is treated elsewhere.

It was mentioned earlier that the frequency of gravitational waves is proportional to the square root of the mean density of the emitting system; this is approximately true for any gravitating system. For example, neutron stars usually have masses of around 1.4 solar masses and radii of the order of 10 km; thus, if we use these numbers in the relation $f \sim \sqrt{GM/R^3}$, we find that an oscillating neutron star will emit gravitational waves primarily at frequencies of 2–3 kHz. By analogy, a black hole a 100 times more massive than the sun will have a radius of $\sim$300 km and the natural oscillation frequency will be around 100 Hz. Finally, for a binary system, Kepler's law (see Section II.E) provides a direct and accurate estimation of the frequency of the emitted gravitational waves. For two 1.4-solar-mass neutron stars orbiting around each other at a distance of 160 km, Kepler's law predicts an orbital frequency of 50 Hz, which leads to an observed gravitational wave frequency of 100 Hz.

A. Radiation from Gravitational Collapse

Type II supernovae are associated with the core collapse of a massive star together with a shock-driven expansion of a luminous shell which leaves behind a rapidly rotating neutron star or, if the core has mass of $>$2–3 solar masses, a black hole. The typical signal from such an explosion is broadband and peaked at around 1 kHz. Detection of such a signal has been the goal of detector development over the last three decades. However, we still know little about the efficiency with which this process produces gravitational waves. For example, an exactly spherical collapse will not produce any gravitational radiation at all. The key issue is the kinetic energy of the nonspherical motions since the gravitational wave amplitude is proportional to this [Eq. (30)]. After 30 years of theoretical and numerical attempts to simulate gravitational collapse, there is still no great progress in understanding the efficiency of this process in producing gravitational waves. For a conservative estimate of the energy in nonspherical motions during the collapse, relation (31) leads to events of an amplitude detectable in our galaxy, even by bar detectors. The next generation of laser interferometers would be able to detect such signals from the Virgo cluster at a rate of a few events per month.

The main source of nonsphericity during the collapse is the angular momentum. During the contraction phase, the angular momentum is conserved and the star spins up to rotational periods of the order of 1 msec. In this case, a number of consequent processes with large luminosity might take place in this newly born neutron star. A number of instabilities, such as the so-called bar mode instability and the r-mode instability, may occur which radiate copious amounts of gravitational radiation immediately after

the initial burst. Gravitational wave signals from these rotationally induced stellar instabilities are detectable from sources in our galaxy and are marginally detectable if the event takes place in the nearby cluster of about 2500 galaxies, the Virgo cluster, 15 Mpc away from the earth. Additionally, there will be weaker but extremely useful signals due to subsequent oscillations of the neutron star; f, p, and w modes are some of the main patterns of oscillations (normal modes) of the neutron star that observers might search for. These modes have been studied in detail, and once detected in the signal, they would provide a sensitive probe of the neutron star structure and its supranuclear equation of state. Detectors with high sensitivity in the kilohertz band will be needed in order to fully develop this so-called gravitational wave asteroseismology.

If the collapsing central core is unable to drive off its surrounding envelope, then the collapse continues and finally a black hole forms. In this case the instabilities and oscillations discussed above are absent and the newly formed black hole radiates away within a few milliseconds any deviations from axisymmetry and ends up as a rotating or *Kerr* black hole. The characteristic oscillations of black holes (normal modes) are well studied, and this unique ringing down of a black hole could be used as a direct probe of their existence. The frequency of the signal is inversely proportional to the black hole mass. For example, it was stated earlier that a 100-solar-mass black hole will oscillate at a frequency of ~100 Hz (an ideal source for LIGO), while a supermassive one with mass 10^7 solar masses, which might be excited by an infalling star, will ring down at a frequency of 10^{-3} Hz (an ideal source for LISA). The analysis of such a signal should reveal directly the two parameters that characterize any (uncharged) black hole, namely its mass and angular momentum.

B. Radiation from Spinning Neutron Stars

A perfectly axisymmetric rotating body does not emit any gravitational radiation. Neutron stars are axisymmetric configurations, but small deviations cannot be ruled out. Irregularities in the crust (perhaps imprinted at the time of crust formation), strains that have built up as the stars have spun down, off-axis magnetic fields, and/or accretion could distort the axisymmetry. A bump that might be created at the surface of a neutron star spinning with frequency f will produce gravitational waves at a frequency of $2f$ and such a neutron star will be a weak but continuous and almost monochromatic source of gravitational waves. The radiated energy comes at the expense of the rotational energy of the star, which leads to a spindown of the star. If gravitational wave emission contributes considerably to the observed spindown of pulsars, then we can estimate the amount of the emitted energy. The corresponding amplitude of gravitational waves from nearby pulsars (a few kpc away) is of the order of $h \sim 10^{-25}$–10^{-26}, which is extremely small. If we accumulate data for sufficiently long time, e.g., 1 month, then the effective amplitude, which increases as the square root of the number of cycles, could easily go up to the order of $h_c \sim 10^{-22}$. We must admit that we are extremely ignorant of the degree of asymmetry in rotating neutron stars, and these estimates are probably very optimistic. On the other hand, if we do not observe gravitational radiation from a given pulsar we can place a constraint on the degree of nonaxisymmetry of the star.

C. Radiation from Binary Systems

Binary systems are the best sources of gravitational waves because they emit copious amounts of gravitational radiation, and for every system we know exactly the amplitude and frequency of the gravitational waves in terms of the masses of the two bodies and their separation (see Section II.E). If a binary system emits detectable gravitational radiation in the bandwidth of our detectors, we can easily identify the parameters of the system. According to the formulas of Section II.E, the observed frequency change will be $\dot{f} \sim f^{11/3} M_{\text{chirp}}^{5/3}$ and the corresponding amplitude will be $h \sim M_{\text{chirp}}^{5/3} f^{2/3}/r = \dot{f}/f^3 r$, where $M_{\text{chirp}}^{5/3} = \mu M^{2/3}$ is a combination of the total and reduced mass of the system called the *chirp mass*. Since both frequency f and its rate of change $\dot{f}$ are measurable quantities, we can immediately compute the chirp mass (from the first relation), thus obtaining a measure of the masses involved. The second relation provides a direct estimate of the distance of the source. These relations have been derived using the Newtonian theory to describe the orbit of the system and the quadrupole formula for the emission of gravitational waves. Post-Newtonian theory—inclusion of the most important relativistic corrections in the description of the orbit—can provide more accurate estimates of the individual masses of the components of the binary system.

When analyzing the data of periodic signals, the *effective amplitude* is not the amplitude of the signal alone, but $h_c = \sqrt{n} \cdot h$, where n is the number of cycles of the signal within the frequency range where the detector is sensitive. A system consisting of two typical neutron stars will be detectable by LIGO when the frequency of the gravitational waves is ~10 Hz until the final coalescence around 1000 Hz. This process will last for about 15 min and the total number of observed cycles will be of the order of 10^4, which leads to an enhancement of the detectability by a factor of 100. Binary neutron star systems and binary black hole systems with masses of the order of 50 solar masses are the primary sources for LIGO. Given the anticipated

sensitivity of LIGO, binary black hole systems are the most promising sources and could be detected as far as 200 Mpc away. The event rate with the present estimated sensitivity of LIGO is probably a few events per year, but future improvement of detector sensitivity (the LIGO II phase) could lead to the detection of at least one event per month. Supermassive black hole systems of a few million solar masses are the primary sources for LISA. These binary systems are rare, but due to the huge amount of energy released, they should be detectable from as far as the boundaries of the observable universe.

D. Cosmological Gravitational Waves

One of the strongest pieces of evidence in favor of the Big Bang scenario is the 2.7 K cosmic microwave background radiation. This thermal radiation first bathed the universe around 1 million years after the Big Bang. By contrast, the gravitational radiation background anticipated by theorists was produced at Planck times, i.e., at 10^{-43} sec or earlier after the Big Bang. Such gravitational waves have traveled almost unimpeded through the universe since they were generated. The observation of cosmological gravitational waves will be one of the most important contributions of gravitational wave astronomy. These primordial gravitational waves will be, in a sense, another source of noise for our detectors and so they will have to be much stronger than any other internal detector noise in order to be detected. Otherwise, confidence in detecting such primordial gravitational waves could be gained by using a system of two detectors and cross-correlating their outputs. The two LIGO detectors are well placed for such a correlation.

SEE ALSO THE FOLLOWING ARTICLES

COSMOLOGY • GLOBAL GRAVITY MODELING • GRAVITATIONAL WAVE ASTRONOMY • NEUTRON STARS • PULSARS • RELATIVITY, GENERAL • SUPERNOVAE

BIBLIOGRAPHY

Blair, D. G. (1991). "The Detection of Gravitational Waves," Cambridge University Press, Cambridge.

Marck, J.-A., and Lasota, J.-P. (eds.). (1997). "Relativistic Gravitation and Gravitational Radiation," Cambridge University Press, Cambridge.

Saulson, P. R. (1994). "Fundamentals of Interferometric Gravitational Wave Detectors," World Scientific, Singapore.

Thorne, K. S. (1987). "Gravitational radiation." *In* "300 Years of Gravitation" (Hawking, S. W., and Israel, W., eds.), Cambridge University Press, Cambridge.



Heat Transfer

George Alanson Greene
Brookhaven National Laboratory

I. Conduction Heat Transfer
II. Heat Transfer by Convection
III. Thermal Radiation Heat Transfer
IV. Boiling Heat Transfer
V. Physical and Transport Properties

GLOSSARY

Boiling The phenomenon of heat transfer from a surface to a liquid with vaporization.
Conduction Heat transfer within a solid or a motionless fluid by transmission of mechanical vibrations and free electrons.
Convection Heat transfer within a flowing fluid by translation of macroscopic fluid volumes from hot regions to colder regions.
Heat transfer coefficient An engineering approximation defined by Newton's law of cooling which relates the heat flux to the overall temperature difference in a system.
Thermal radiation The transport of thermal energy from a surface by nonionizing electromagnetic waves.
Temperature A scalar quantity which defines the internal energy of matter.

HEAT TRANSFER plays an essential role in everything that we do. Our bodies are exposed to a changing environment yet to live we must remain at 98.6°F. The dynamics of our planet's atmosphere and oceans are driven by the seasonal variations in heat flux from our Sun. These dynamics, in turn, dictate whether it will rain or snow, whether there will be hurricanes or tornadoes, drought or floods, if crops will grow or die. We burn fuels to heat our homes, our power plants create steam to turn turbines to make electricity, and we reverse the process to cool our dwellings with air conditioning for comfort. There are several modes for the transport of heat which we experience daily. Among these are conduction of heat in solids by molecular vibrations, convection of heat in fluids by the motion of fluid elements from hot to cold regions, thermal radiation in which heat is transferred from surface to surface by electromagnetic radiation, and boiling heat transfer in which heat is transferred from a surface by causing a liquid-to-vapor phase change in an adjacent fluid.

I. CONDUCTION HEAT TRANSFER

Heat transfer in opaque solids occurs exclusively by the process of conduction. If a solid (or a stationary liquid or gas) is transparent or translucent, heat can be transferred in a solid by both conduction and radiation, and in fluids, heat can be transferred by conduction, radiation, and

convection. In general, materials in which heat is transferred by conduction only are solid bodies; the addition of convection and radiation to these systems enters the solutions through conditions imposed at the boundaries.

A. Physics of Conduction and Thermal Conductivity

The mechanisms of heat conduction depend to a great extent on the structure of the solid. For metals and other electrically conducting materials, heat is conducted through a solid by atomic and molecular vibrations about their equilibrium positions and by the mobility of free conduction-band electrons through the solid, the same electrons which conduct electricity. There is, in fact, a rigorous relationship between the thermal and the electrical conductivities in metals, known as the Franz–Wiedemann law. In nonmetallic or dielectric materials, lattice vibrations induced by atomic vibrations, otherwise known as phonons, are the principle mechanism of heat conduction. A phonon can be considered a quantum of thermal energy in a thermoelastic wave of fixed frequency passing through a solid, much like a photon is a quantum of energy of a fixed frequency in electromagnetic radiation theory, hence the origin of the name. It is the absence of free conduction-band electrons that make dielectrics poor heat and electrical conductors, relying only on phonons or lattice vibrations to transfer heat energy through a solid. This is intuitively less efficient than conduction in a metal or conductor as will be discussed. It is also clear that in dielectrics which rely on phonon transport for heat transfer, anything which reduces the phonon transport in a material will correspondingly reduce its heat transfer efficiency. An example is the effect of dislocations or impurities in crystals and alloying in metals.

The roots of our understanding of thermal conductivity by phonon transport come from kinetic theory and particle physics. Phonon transport in a dielectric is analogous to the thermal conductivity of a gas which depends on collisions between gas molecules to transfer heat. Considering the phonons as particles in the spirit of the dual particle-wave nature of electromagnetic theory, the thermal conductivity of a dielectric solid can be shown to be represented by the relationship given as $k_p = \rho c_v v \lambda/3$, where ρ is the phonon density, c_v is the heat capacity at constant volume, v is the average phonon velocity, and λ is the mean free path of the phonon. For heat conduction by phonon transport, the phonon velocity is on the order of the sound speed in the solid and the mean free path is on the order of the interatomic spacing. Although the phonon density increases with increasing temperature, the thermal conductivity may remain unchanged or even decrease as the temperature increases if the effect of the vibrations is to diminish the mean free path by an equivalent factor or more. If a dielectric is raised to a very high temperature, heat conduction is increased by thermal excitation of bound electrons which causes them to take on the characteristics of free electrons as in metals, hence the increased thermal conductivity as we find in metals. In extreme cases, this can be accompanied by electron or x-ray emission from the solid.

In metals, conduction by phonons is enhanced by conduction by electrons, as just described for high-temperature dielectrics. The derivation from quantum mechanics is parallel to that for phonon transport, except that c is the electron heat capacity, v is the Fermi velocity of the free electrons, and λ is the electronic mean free path of the valence electrons. Due to its complexity, the details of the derivation of the electron contribution to the thermal conductivity will not be presented. However, it is easy to show that in a metal, the total thermal conductivity is the sum of the phonon contribution and the electron contribution, $k = k_p + k_e$. In pure metals, the electron contribution to the thermal conductivity may be 30 times greater than the phonon contribution at room temperature. For a more rigorous derivation of the thermal conductivity, the reader is directed to the literature on solid-state physics and quantum mechanics.

B. Fundamental Law of Heat Conduction

The second law of thermodynamics requires that heat is transferred from one body to another body only if the two bodies are at different temperatures, and that the heat flows from the body at the highest temperature to the body at the lowest temperature. In essence, this is a statement that a thermal gradient must exist in the solid and that heat flows down the thermal gradient. In addition, the first law of thermodynamics requires that the thermal energy is conserved in the absence of heat sources or sinks in the body. It follows from this that a body has a temperature distribution which is a function of space and time, $T = T(x, y, z, t)$, and that the thermal field within the solid is constructed by the superposition of an infinite number of isothermal surfaces which never intersect, lest some point of intersection in space be simultaneously at two or more temperatures which is impossible.

Consider a semi-infinite solid whose boundaries are parallel and isothermal at different temperatures. Eventually the temperature distribution within the body will become invariant with time and the heat flow from surface one to surface two becomes $q_{1-2} = -kA(T_1 - T_2)/d$, where q_{1-2} is the heat flux, A is the area normal to the heat flux, T_1 and T_2 are the temperatures of the two isothermal bounding surfaces and d is the separation between the surfaces.

In the limit that the coordinate normal to the isothermal plane approaches zero, this equation then becomes

$$q_n = -kA\frac{\partial T}{\partial n} \tag{1a}$$

or

$$q_n'' = -k\frac{\partial T}{\partial n}, \tag{1b}$$

which is known as Fourier's heat conduction equation. The heat flow per unit area per unit time across a surface is called the heat flux q'' and has units of W/m^2. The heat flux is a vector and can be calculated for any point in a solid if the temperature field and thermal conductivity are known.

C. Differential Heat Conduction Equation

The differential heat conduction equations derive from the application of Fourier's law of heat conduction, and the basic character of these equations is dependent upon shape and varies as a function of the coordinate system chosen to represent the solid. If Fourier's equation is applied to a simple, isotropic solid in Cartesian coordinates and if the thermal conductivity is assumed to be constant, the equation for the transient conservation of thermal energy due to conduction of heat in a solid with a heat source (or heat sink) can be derived as follows,

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} + \frac{q'''}{k} = \frac{1}{\alpha}\frac{\partial T}{\partial t}, \tag{2}$$

where q''' is the volumetric heat source and α is the thermal diffusivity, $\alpha = k/\rho c$. If the heat source is equal to zero, this reduces to the Fourier equation. If the temperature in the solid is invariant with respect to time, this becomes the Poisson equation. Furthermore, if the temperature is time-invariant and the heat source is zero, this becomes the Laplace equation. Other forms of the thermal energy conservation equations in a solid can be derived for other coordinate systems and the eventual solutions depend on the initial and boundary conditions which are imposed. It is the solutions to these equations as given previously that is modeled in commercially available computer analysis packages for heat transfer solutions in solids. The reader is referred to VanSant for a thorough listing of analytical solutions to the heat conduction equations in many coordinate systems and subject to numerous initial and boundary conditions in order to experience the elegance of analytical solutions to problems in heat transfer physics and to appreciate the relationship between heat transfer and mathematics.

A curiosity of the parabolic differential form of the heat conduction equation just presented is that it implies that the velocity of propagation of a thermal wave in a solid is infinite. This is a consequence of the fact that the solution predicts that the effects of a thermal disturbance in a solid are felt immediately at a distance infinitely removed from the disturbance itself. This is in spite of the definition of the thermal conductivity which is based upon finite speed of propagation of free electrons or phonons in matter. In practical applications, this outcome is inconsequential because the effect at infinity is generally small. However, there are circumstances in which this peculiarity in the equations may actually become significant and lead to erroneous results, for instance, in heat transfer problems at very low temperatures or very short time scales, in which cases the finite speed of propagation of heat becomes important. Two examples of such circumstances which can be encountered in practice are cryogenic heat transfer near absolute zero and rapid energy transfer in materials due to subatomic particles which travel at the speed of light. It has been suggested that the form of the differential equations for conduction heat transfer should be the damped-wave or the hyperbolic heat conduction equation, often called the telegraph equation, which includes the finite speed of propagation of heat, C, as shown below without derivation.

$$\frac{1}{C^2}\frac{\partial^2 T}{\partial t^2} + \frac{1}{\alpha}\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}. \tag{3}$$

For most practical problems in heat conduction, the solutions to the parabolic and hyperbolic heat conduction equations are essentially identical; however, the cautions offered in Eq. (3) should be evaluated in circumstances where the finite propagation speed could become important, especially when using commercial equation solvers which will undoubtedly not model the hyperbolic effect just described. Rendering the damped-wave equation dimensionless will reveal to the analyst when the wave propagation term and the diffusion term on the LHS of the hyperbolic heat conduction equation are of the same order of magnitude and both must be included in the solution, for instance, when $\Delta t \sim (\alpha/C^2)$.

A continued discussion of conduction heat transfer in solids would require the solutions to many special heat transfer cases for which the parabolic heat conduction equation can be easily integrated. Examples of these can be found in every heat transfer text book, and they will not be solved here. However, two examples will be discussed which illustrate powerful techniques for solving contemporary heat transfer problems. These are the lumped heat capacity approximation for transient heat conduction in solids with convection, and the numerical decomposition of the differential heat conduction equation for finite difference computer analysis. By necessity, these discussions will be brief but illustrative.

D. Lumped Heat Capacity Approximation in Transient Conduction

Some heat transfer systems, usually involving a "small" body or a body which is thin in the direction of heat transfer, can be analyzed under the assumption that their temperature is uniform spatially, only a function of time. This is called the lumped heat capacity assumption. The system can be analyzed as a function of time only, greatly simplifying the analysis. This situation can be illustrated by considering a small, spherical object at an initial temperature T_0 which is suddenly submerged in a fluid at temperature T_∞ which imposes a heat transfer coefficient at the surface of the sphere h with the units W/m^{2}K. If the sphere has density ρ , specific heat c_p, surface area A, and volume V, the transient energy conservation equation can be written as follows:

$$\frac{dT(t)}{dt} = -\frac{hA}{\rho c_p V}(T(t) - T_\infty), \tag{4}$$

and the solution for the time-dependent dimensionless temperature of the sphere θ becomes as follows, $\theta = \exp(-\tau) = \exp(-\mathrm{Bi}\cdot\mathrm{Fo})$, where Bi is the Biot number ($\mathrm{Bi} = h\ell/k$) which is the ratio of the internal heat transfer resistance to the external heat transfer resistance, and Fo is the Fourier number ($\mathrm{Fo} = \alpha t/\ell^2$), the dimensionless time. In order to simplify the solution for the transiently cooled sphere, a condition was imposed that the spatial variations of the temperature in the sphere were small. This condition is satisfied if the resistance to heat transfer inside the object is small compared to the external resistance to heat transfer from the sphere to the fluid. Mathematically, this is stated that the Biot number ≤ 0.1, a factor of an order of magnitude. If this condition is satisfied, heat transfer solutions can be greatly simplified.

E. Finite Difference Representation of Steady-State Heat Conduction

In practice, it is frequently not possible to achieve analytical solutions to heat transfer problems in spite of simplifications and approximations. It is often necessary to resort to numerical solutions because of complexities involving geometry and shape, variable physical and transport properties, and complex and variable initial and boundary conditions. Figure 1 shows a rectilinear two-dimensional solid, divided into a grid of equally spaced nodes; it will be assumed that a steady-state temperature field exits. Three nodes are depicted in Fig. 1: (a) an interior node which is surrounded by other nodes in the solid, (b) a node on the insulated boundary of the solid, and (c) a node on the convective boundary of the solid.

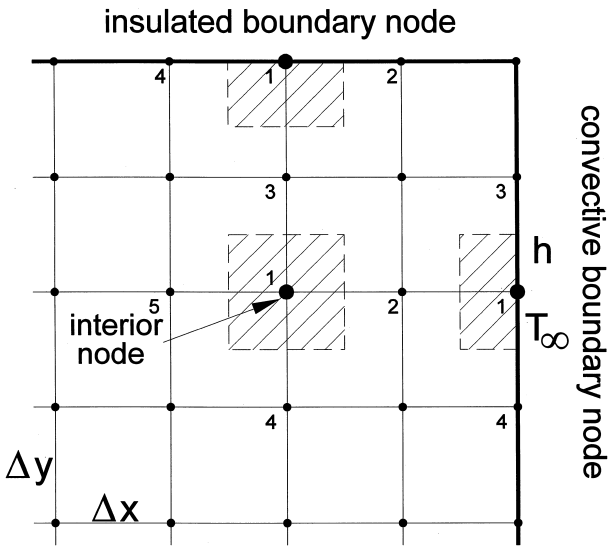


FIGURE 1 Cartesian coordinate grid for finite difference analysis.

An analytical solution to such a simple heat transfer problem could be a formidable task; however, analysis by decomposing the energy equation into a form suitable for finite difference numerical analysis will greatly simplify the task. In other words, an energy balance is performed on each shaded control volume (one for every node), allowing for heat transfer across each face of the control volume from surrounding nodes or, in the case of the convective boundary, the surrounding fluid. Assuming for convenience that $\Delta x = \Delta y$, the temperature is not a function of time and the thermal conductivity is a constant, the finite difference equations can be easily derived, and they are presented here for the three nodes in the example in Fig. 1.

$$\begin{aligned} &\text{Interior node:} && 0 = T_2 + T_3 + T_4 + T_5 - 4T_1 \\ &\text{Insulated boundary:} && 0 = T_2 + T_4 + 2T_3 - 4T_1 \end{aligned} \tag{5}$$

Convective boundary:

$$0 = \frac{1}{2}(2T_3 + T_2 + T_4) + \left(\frac{h\cdot\Delta x}{k}\right)T_\infty - \left(\frac{h\cdot\Delta x}{k} + 2\right)T_1.$$

Note the appearance in the convective boundary node equation of the term $(h\Delta x/k)$, the finite difference form of the Biot number which was introduced in the preceding section. Such equations can be written for all the nodes and assembled in a manner convenient for iterative solution. Simple examples such as those shown here will rapidly converge; more complex problems will require more complex algorithms and stringent convergence criteria.

II. HEAT TRANSFER BY CONVECTION

In the preceding section, we discussed the mechanisms of conduction heat transfer as the sole agent which transports heat energy within a solid. Convection was only considered insofar as it entered the problem through the boundary conditions. For fluids, however, this is true only under the conditions that the fluid is motionless, a condition almost never realized in practice. In general, fluids are in motion either by pumping or by buoyancy, and the heat transfer in fluids in motion is enhanced over conduction because the moving fluid particles carry heat with them as internal energy; the transport of heat through a fluid by the motion of macroscopic fluid particles is called convection. We will now consider methods of modeling convective heat transfer and the concept of the heat transfer coefficient which is the fundamental variable in convection. The analysis of convective heat transfer is more complex than conduction in solids because the motion of the fluid must be considered simultaneously with the energy transfer process. The general approach assumes that the fluid is a continuum instead of the more basic and complex approach assuming individual particles. Although fundamental issues such as the thermodynamic state and the transport properties of the fluid cannot be solved theoretically by the continuum approach, the solutions to the fluid mechanics and heat transfer are made more tractable; parallel studies at the molecular level can resolve the thermodynamic and transport issues. In practice, the thermodynamic and transport properties, although available from theoretical studies on a molecular level, are generally input to the study of heat transfer empirically.

A. Internal and External Convective Flows

There are two general classes of problems in convective heat transfer: internal convection in channels and pipes in which the flow patterns become fully developed and spatially invariant after traversing an initial entrance length and the heat flux is uniform along the downstream surfaces, and external convection over surfaces which produces a shear or boundary layer which continues to grow in the direction of the flow and which never becomes fully developed or spatially invariant. Both internal duct flows and external boundary layer flows can be either laminar or turbulent, depending upon the magnitude of a dimensionless parameter of the fluid mechanics known as the Reynolds number. For internal flows, the flow is laminar if the Reynolds number is less than 2×10^3; for external flows, the rule of thumb is that the flow is laminar if the Reynolds number is less than 5×10^5. Schematic representations of both an internal flow case and an external boundary layer are shown in Figs. 2a,b, respectively.

B. Fluid Mechanics and the Reynolds Number

Steady flow in a channel (internal flow) and external flow over a boundary are governed by a balance of forces in the fluid in which inertial forces and pressure forces are balanced by viscous forces on the fluid. This leads to the familiar concept of a constant pressure drop in a water pipe which provides the force to overcome friction along the pipe walls and thus provides the desired flow rate of water out the other end. For Newtonian fluids, viscous or shear forces in the fluid are described by a relationship between the stress (force/unit area) between the fluid layers which

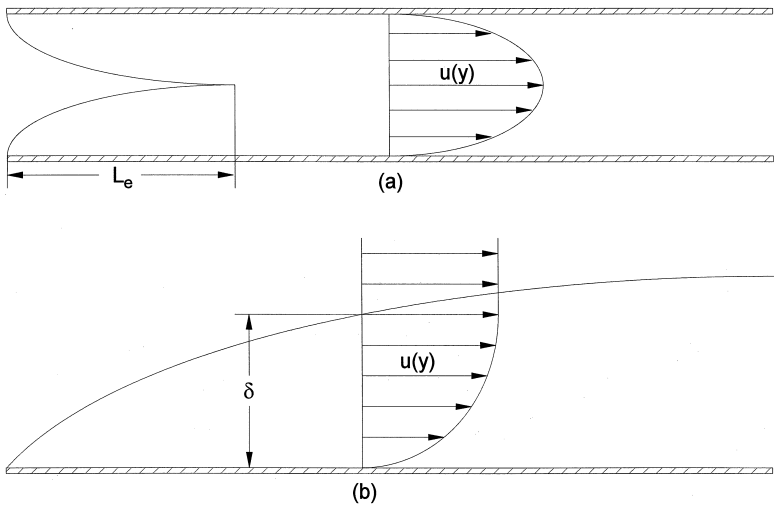


FIGURE 2 Schematic of internal flow and external flow boundary layers.

results in a shear of the velocity field in the fluid as follows, $\tau = \mu \cdot \partial u/\partial y$, where τ is the shear stress in the fluid, $\partial u/\partial y$ is the rate of strain of the fluid, and μ, the constant of proportionality, is a fluid transport property known as the dynamic viscosity. This is the Newtonian stress–strain relationship and it forms the basis for the fundamental equations of fluid mechanics. The force balance in the direction of flow which provides for a state of equilibrium on a fluid element in the flow can be written as a balance of differential pressure forces normal to the fluid element by tangential shear forces on the fluid element as shown in the following:

$$-\frac{\partial P}{\partial x} + \frac{\partial \tau}{\partial y} = \sum F_x = \frac{D}{Dt}(\rho u). \tag{6}$$

Substituting the Newtonian stress–strain relationship into this force balance, we find that

$$-\frac{\partial P}{\partial x} + \mu \frac{\partial^2 u}{\partial y^2} = \rho\left[u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y}\right] \tag{7a}$$

and in dimensionless form, this becomes

$$-\frac{\partial P^*}{\partial x^*} + \left(\frac{\mu}{\rho UL}\right)\frac{\partial^2 u^*}{\partial y^{*2}} = u^*\frac{\partial u^*}{\partial x^*} + v^*\frac{\partial u^*}{\partial y^*}, \tag{7b}$$

where the quantity $(\rho UL/\mu)$ is called the Reynolds number of the flow, and it represents the ratio of inertial forces to viscous forces in the fluid. The Reynolds number, sometimes written as $\mathrm{Re} = UL/\nu$, where ν is the kinematic viscosity, $\nu = \mu/\rho$, is the similarity parameter of fluid mechanics which provides the convenience of similarity solutions to general classes of fluid mechanics problems (i.e., pressure drop in laminar or turbulent pipe flow can be scaled by the Reynolds number, regardless of the velocity, diameter, or viscosity) and is the parameter which predicts when laminar conditions transition to turbulence. The Reynolds number plays a fundamental role in predicting convection and convective heat transfer.

C. The Convective Thermal Energy Equation and the Nusselt Number

In order to solve for heat transfer in convective flows, an energy balance is constructed on an elemental fluid element. In Cartesian coordinates, this usually involves the balance of convection of heat into and out of the elemental volume in the direction of the flow (x-direction) and conduction of heat into and out of the elemental volume transverse to the flow (y-direction). Taylor series expansions of the convective heat flux in the x-direction and the conduction heat flux in the y-direction permit the representation of the heat balance on the differential fluid volume in differential form. The statement of thermal energy conservation on the differential unit volume of fluid $dx \cdot dy$ can be written in the form of the laminar boundary layer equation as

$$\rho c_p\left[u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y}\right] = k\frac{\partial^2 T}{\partial y^2}. \tag{8}$$

Furthermore, if the velocity, temperature, and coordinates are nondimensionalized by the characteristic scales of the problem, such as the maximum velocity U, the overall length L, and the overall temperature difference $T_\mathrm{w} - T_\infty$, the dimensionless form of the laminar boundary layer equation becomes

$$u^*\frac{\partial \theta}{\partial x^*} + v^*\frac{\partial \theta}{\partial y^*} = \frac{1}{\mathrm{Re}\cdot\mathrm{Pr}}\frac{\partial^2 \theta}{\partial y^{*2}}, \tag{9}$$

subject to the appropriate boundary conditions. Note the appearance in Eq. (9) of the familiar Reynolds number, $\mathrm{Re} = UL/\nu$, and the appearance of another dimensionless parameter, the Prandtl number, $\mathrm{Pr} = \mu\cdot c/k$, which includes the physical and transport properties of fluid mechanics and heat transfer. In simple terms, the Prandtl number represents the ratio of the thickness of the hydrodynamic boundary layer to the thickness of the thermal boundary layer. If the Prandtl number equals unity, both boundary layers grow at the same rate. For all problems of convective heat transfer in fluids, the dominant dimensionless scaling or modeling parameters are the Reynolds number and the Prandtl number. Solutions to the thermal energy equations of convective heat transfer are complex and can only be described here in general terms. To express the overall effect of convection on heat transfer, we call upon Newton's law of cooling given by

$$q = hA(T_w - T_\infty) = -kA\left(\frac{\partial T}{\partial y}\right)_w, \tag{10a}$$

where h is the heat transfer coefficient which has units of $(\mathrm{W/m^2K})$. The heat transfer coefficient can also be written as

$$h = \frac{-k(\partial T/\partial y)_w}{(T_w - T_\infty)}. \tag{10b}$$

The dimensional solutions to the laminar boundary layer equations (and the thermal convective energy equations in general) involve solutions as shown previously for the heat transfer coefficient. The solution for the heat transfer coefficient may be nondimensionalized by multiplying by the characteristic length scale of the problem and dividing by the thermal conductivity. In this manner, the dimensionless heat transfer coefficient is introduced, and is called the Nusselt number,

$$Nu = \frac{h\cdot x}{k} = f(\mathrm{Re}, \mathrm{Pr}), \tag{11}$$

which is the general form of most solutions in forced-flow convective heat transfer.

In closing the discussion on convective heat transfer, it would be useful to present two examples of the dimensionless heat transfer coefficient, the solution to the thermal energy equation, for the two cases presented earlier in Fig. 2. The first example is the laminar flow external boundary layer which was depicted in Fig. 2b. For the case of laminar boundary convective heat transfer over a horizontal, flat surface, the dimensionless heat transfer coefficient, which is the solution to the thermal energy equation becomes as follows:

$$Nu(x) = \frac{h(x) \cdot x}{k} = 0.332 \, \mathrm{Pr}^{1/3} \, \mathrm{Re}(x)^{1/2}. \qquad (12)$$

The second example is fully-developed flow in a smooth circular pipe which was depicted in Fig. 2a. For the case of fully developed turbulent flow in a smooth circular pipe, the solution of the convective thermal energy equation becomes as follows:

$$Nu_d = \frac{h \cdot d}{k} = 0.023 \, \mathrm{Re}(d)^{0.8} \, \mathrm{Pr}^{\mathrm{n}}, \qquad (13)$$

where $n = 0.4$ for heating and $n = 0.3$ for cooling. Equation (13) is called the Dittus–Boelter equation for turbulent heat transfer in a pipe. The derivations given in this chapter were, by necessity, simplifications of more rigorous derivations which may be found in the bibliography.

III. THERMAL RADIATION HEAT TRANSFER

In the preceding sections, we examined two fundamental modes of heat transfer, conduction and convection, and have shown how they are developed from a fundamental theoretical approach. We now turn our attention to the third fundamental mode of heat transfer, thermal radiation.

A. Physical Mechanisms of Thermal Radiation

Thermal radiation is the form of electromagnetic radiation that is emitted by a body as a result of its temperature. There are many types of electromagnetic radiation, some is ionizing and some is nonionizing. Electromagnetic radiation generally becomes more ionizing with increasing frequency, for instance x-rays and γ-rays. At lower frequencies, electromagnetic radiation becomes less ionizing, for instance, visible, thermal, and radio wave radiation. However, this is not a hard and fast rule. The spectrum of thermal radiation includes the portion of the frequency band of the electromagnetic spectrum which includes infrared, visible, and ultraviolet radiation. Regardless of the type of electromagnetic radiation being considered, all electromagnetic radiation is propagated at the speed of light, $c = 3 \times 10^{10}$ cm/s, and this speed is equal to the product of the wavelength and frequency of the radiation, $c = \lambda\nu$, where λ is the wavelength and is ν the frequency. The portion of the electromagnetic spectrum which is considered thermal covers the range of wavelength from 0.1 to 100 μm; in comparison, the visible light portion of the thermal spectrum in which humans can see is very narrow, covering only from 0.35 to 0.75 μm. If we were insects, we might see in the infrared range of the spectrum. If we did, warm bodies would look like multicolored objects but the glass windows in our homes would be opaque because infrared is reflected by glass just like visible light is reflected by mirrors. Since the windows in our homes are transparent to visible light, they let in solar radiation which is emitted in the visible spectrum; since they are opaque to infrared radiation, the surfaces in your house which radiate in the infrared do not radiate out to space at night; your house loses heat by conduction through the walls and convection from the outside surfaces.

The emission or propagation of thermal energy takes place as discrete photons, each having a quantum of energy E given by $E = h\nu$, where h is Plank's constant ($h = 6.625 \times 10^{-34}$ J $\cdot$ s). An analogy is sometimes used to characterize the propagation of thermal radiation as particles such as the molecules of a gas, each having mass, momentum, and energy, a so-called photon "gas." In this fashion, we have that the energy of the photons in the "photon gas" is $E = h\nu = mc^2$, the photon mass is $m = h\nu/c^2$, and the photon momentum is $p = h\nu/c$. It follows from statistical thermodynamics that the radiation energy density per unit volume per unit wavelength can be derived as

$$u_\lambda = \frac{3\pi hc\lambda^{-5}}{\exp(hc/\lambda kT) - 1}, \qquad (14)$$

where k is Boltzmann's constant ($k = 1.38 \times 10^{-23}$ J/molecule $\cdot$ K). If the energy density of the radiating gas is integrated over all wavelengths, the total energy emitted is proportional to the absolute temperature of the emitting surface to the fourth power as

$$E_b = \sigma T^4, \qquad (15)$$

where E_b is the energy radiated by an ideal radiator (or black body) and σ is the Stefan–Boltzmann constant ($\sigma = 5.67 \times 10^{-8}$ W/m^{2}K^4). E_b is called the emissive power of a black body. The term "black body" should be taken with caution for although most surfaces which look black to the eye are ideal radiators, other surfaces such as ice and some white paints are also black at long wavelengths. Equation (15) is known as the Stefan–Boltzmann law of radiation heat transfer for an ideal thermal radiator. We have now developed three fundamental laws of classical heat transfer:

- Fourier's law of conduction heat transfer
- Newton's law of convective cooling
- Stefan–Boltzmann law of thermal radiation

B. Radiation Properties

The properties of thermal radiation are not dissimilar to our experience with visible light. When thermal radiation is incident upon a surface, part is reflected, part is absorbed, and part is transmitted. The fraction reflected is called the reflectivity ρ, the fraction absorbed is called the absorptivity α, and the fraction transmitted is called the transmissivity τ. These three variables satisfy the identity that $\alpha+\rho+\tau=1$. Since most solid bodies do not transmit thermal radiation, $\tau=0$ and the identity reduces to $\alpha+\rho=1$.

There are two types of surfaces when it comes to the reflection of thermal radiation from a surface: specular and diffuse. If the angle of incidence of incoming radiation is equal to the angle of reflected radiation, the reflected radiation is called specular. If the reflected radiation is distributed uniformly in all directions regardless of the angle of incidence, the reflected radiation is called diffuse. In general, polished smooth surfaces are more specular and rough surfaces are more diffuse. The emissive power of a surface E is defined as the energy emitted from the surface per unit area per unit time. If you consider a body of surface area A inside a black enclosure and in thermal equilibrium with the enclosure, an energy balance on the enclosed surface states that $E\cdot A=q_i\cdot A\cdot\alpha$; in other words, the energy emitted from the body is equal to the fraction of the incident energy absorbed from the black enclosure. If the surface inside the black enclosure is itself a black body, the statement of thermal equilibrium then becomes as follows: $E_b\cdot A=q_i\cdot A\cdot(1)$, where $\alpha=1$ for the enclosed black body. Dividing these two statements of thermal equilibrium, we get, $E/E_b=\alpha$; in other words, the ratio of the emissive power of a body to the emissive power of a black body at the same temperature is equal to the absorptivity of the surface, α. If this ratio holds such that the absorptivity α is equal to the emissivity $\in$ for all wavelengths, we have Kirchhoff's law, $\in=\alpha$, for a grey body or for grey body radiation. In other words, the surface is a grey body such that the monochromatic emissivity of the surface $\in_\lambda$ is a constant for all wavelengths.

In practice, the emissivities of various surfaces can vary by a great deal as a function of wavelength, temperature, and surface conditions. A graphical example of the variations in the total hemispherical emissivity of Inconel 718 as a function of surface condition and temperature as reported by the author is shown in Fig. 3. However,

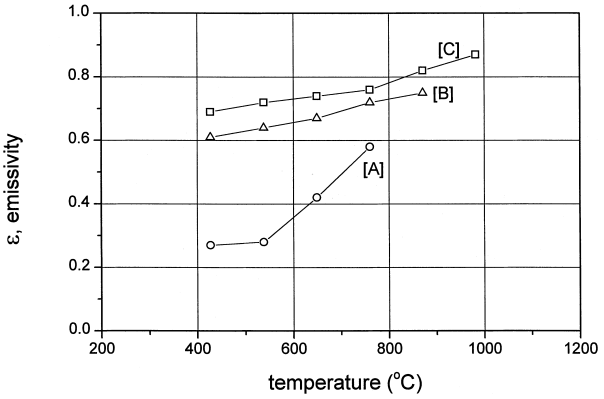


FIGURE 3 Total hemispherical emissivity of Inconel 718: (a) shiny, (b) oxidized in air for 15 min at 815°C, (c) sandblasted and oxidized in air for 15 min at 815°C.

the convenience of the assumptions of a grey body, one whose monochromatic emissivity $\in_\lambda$ is independent of wavelength, and Kirchhoff's law, that $\in=\alpha$, make many practical problems more tractable to solution.

Plank has developed a formula for the monochromatic emissive power of a black body from quantum mechanics as shown in the following:

$$E_{b\lambda}=\frac{C_1\lambda^{-5}}{\exp(C_2/\lambda T)-1}, \tag{16}$$

where $C_1=3.743\times10^8\,\mathrm{W}\cdot\mu\mathrm{m}^4/\mathrm{m}^2$ and $C_2=1.439\times10^4\,\mu\mathrm{m}\cdot\mathrm{K}$, and λ is the wavelength in micrometers. Plank's distribution function given in Eq. (16) predicts that the maximum of the monochromatic black body emissive power shifts to shorter wavelengths as the absolute temperature increases, and that the peak in $E_{b\lambda}$ increases as the wavelength decreases. An illustrative example of the trends of Plank's distribution function is that while a very hot object such as a white-hot ingot of steel radiates in the visible spectrum, $\lambda\sim1\ \mu$m, as it cools it will radiate in increasingly longer wavelengths until it is in the infrared spectrum, $\lambda\sim100\ \mu$m.

A relationship between the temperature and the peak wavelength of Plank's black body emissive power distribution function known as Wein's displacement law is given here:

$$\lambda_{\max}\cdot T=2897.6\ \mu\mathrm{m}\cdot\mathrm{K}. \tag{17}$$

This relationship determines the peak wavelength of the emissive power distribution for a black body at any temperature T. If the body is grey with an average emissivity $\in$, the value of $E_{b\lambda}$ is simply multiplied by $\in$ to get E_λ, but as a first approximation, the peak of Plank's distribution function remains unchanged.

C. Radiation Shape Factors

Surfaces that radiate thermal energy radiate to each other, and it is necessary to know how much heat leaving Surface 1 gets to Surface 2 and vice versa, in order to determine the surface heat flux. The function that determines the amount of heat leaving Surface 1 that is incident on Surface 2 is called the radiation shape factor, F_{1-2}. Consider two black surfaces A_1 and A_2 at two different temperatures T_1 and T_2. The energy leaving A_1 arriving at A_2 is $E_{b1}A_1F_{1-2}$ and the energy leaving A_2 arriving at A_1 is $E_{b2}A_2F_{2-1}$. Since the surfaces are black and all incident energy is absorbed ($\in_1 = \in_2 = 1$), the net radiative energy exchange between the two surfaces is $q_{1-2} = E_{b1}A_1F_{1-2} - E_{b2}A_2F_{2-1}$. Setting both surfaces to the same temperature forces q_{1-2} to zero and $E_{b1} = E_{b2}$, therefore $A_1F_{1-2} = A_2F_{2-1}$. This relationship is known as the reciprocity relationship for radiation shape factors and can be written in general as $A_mF_{m,n} = A_nF_{n,m}$. This relationship is geometrical and applies for grey diffuse surfaces as well as black surfaces. Since our surfaces were black, we can substitute the black body emissive power, $E_{bi} = \sigma T_i^4$, to get the result

$$q_{1-2} = A_1F_{1-2}\sigma\left(T_1^4 - T_2^4\right). \tag{18}$$

In general, the solution for shape factors involves geometrical calculus. However, many shape factors have been tabulated in books which simplify the analyses significantly. There are relations between shape factors which are useful for constructing complex shape factors from an assembly of more simple shape factors. Considerable time could be spent in this discussion; however, these will not be discussed here and the reader is directed to the references for more details.

D. Heat Exchange Between Nonblack Bodies

We have just derived a useful equation for the heat flux between two black, diffuse surfaces $q_{1-2} = A_1F_{1-2}(E_{b1} - E_{b2})$. In analogy to Ohm's law, this can be rewritten in the form of a resistance to heat transfer as $q_{1-2} = (E_{b1} - E_{b2})/R_{\text{spatial}}$ where $R_{\text{spatial}} = 1/(A_1F_{1-2})$. It is implied in this formulation that since both bodies are black and thus perfect emitters, they have no surface resistance to radiation, only a geometrical spatial resistance. If both surfaces were grey, they would have $\in \neq 1$, and there would be associated with each surface a resistance due to the emissivity of each surface, a thermodynamic resistance in addition to the spatial resistance just shown. Let us examine this problem in more general terms. The problem of determining the radiation heat transfer between black surfaces becomes one of determining the geometric shape factor. The problem becomes more complex when considering nonblack bodies because not all energy incident on a surface is absorbed, some is reflected back and some is reflected out of the system entirely. In order to solve the general problem of radiation heat transfer between grey, diffuse, isothermal surfaces, we must define two new concepts, the radiosity J and irradiation G.

The radiosity J is defined as the total radiation which leaves a surface per unit area per unit time, and the irradiation G is defined as the total energy incident on a surface per unit area per unit time. Both are assumed uniform over a surface for convenience. Assuming that $\tau = 0$ and $\rho = (1 - \in)$, the equation for the radiosity J is as follows:

$$J = \varepsilon E_b + \rho G = \varepsilon E_b + (1 - \varepsilon)G. \tag{19}$$

Since the net energy leaving the surface is the difference between the radiosity and the irradiation, we find,

$$\frac{q}{A} = J - G = \varepsilon E_b + (1 - \varepsilon)G - G, \tag{20}$$

and solving for G from Eq. (19) and substituting in Eq. (20), we get the following solution for the surface heat flux:

$$q = \frac{\varepsilon A}{(1 - \varepsilon)}(E_b - J) = \frac{E_b - J}{(1 - \varepsilon)/\varepsilon A}. \tag{21}$$

In another analogy to Ohm's law, the LHS of Eq. (21) can be considered a current, the RHS-top a potential difference, and the RHS-bottom a surface resistance to radiative heat transfer. We now consider the exchange of radiant energy between two surfaces A_1 and A_2. The energy leaving A_1 which reaches A_2 is $J_1A_1F_{1-2}$, and the energy leaving A_2 which reaches A_1 is $J_2A_2F_{2-1}$. Therefore, the net energy transfer from A_1 to A_2 is $q_{1-2} = J_1A_1F_{1-2} - J_2A_2F_{2-1}$, and using the reciprocity relation for shape factors we find,

$$q_{1-2} = (J_1 - J_2)A_1F_{1-2} = \frac{(J_1 - J_2)}{(1/A_1F_{1-2})}, \tag{22}$$

where $(1/A_1F_{1-2})$ is the spatial resistance to radiative heat transfer between A_1 and A_2.

A resistance network may now be constructed for two isothermal, grey, diffuse surfaces in radiative exchange with each other by dividing the overall potential difference by the sum of the three resistances as follows:

$$q_{1-2} = \frac{E_{b1} - E_{b2}}{(1 - \varepsilon_1)/\varepsilon_1A_1 + 1/A_1F_{1-2} + (1 - \varepsilon_2)/\varepsilon_2A_2}$$

$$= \frac{\sigma\left(T_1^4 - T_2^4\right)}{(1 - \varepsilon_1)/\varepsilon_1A_1 + 1/A_1F_{1-2} + (1 - \varepsilon_2)/\varepsilon_2A_2}. \tag{23}$$

This approach can be readily extended to include more than two surfaces exchanging radiant energy but the

equations quickly become unwieldy so no example will be presented here.

One example of a problem which may be easily solved with this network method and which frequently arises in practice, such as in the design of experiments, is the problem of two grey, diffuse, and isothermal infinite parallel surfaces. In this problem, $A_1 = A_2$ and $F_{1-2} = 1$ since all the radiation leaving one surface reaches the other surface. Substituting for F_{1-2} in Eq. (23) and dividing by $A = A_1 = A_2$, we find that the net heat flow per unit area becomes

$$\frac{q_{1-2}}{A} = \frac{\sigma\left(T_1^4 - T_2^4\right)}{1/\varepsilon_1 + 1/\varepsilon_2 - 1}. \tag{24}$$

A second example which serves to illustrate this technique is the problem of two long concentric cylinders, with A_1 being the inner cylinder and A_2 the outer cylinder. Once again, applying Eq. (23) noting that $F_{1-2} = 1$, we find that,

$$\frac{q_{1-2}}{A_1} = \frac{\sigma\left(T_1^4 - T_2^4\right)}{1/\varepsilon_1 + (A_1 / A_2)(1/\varepsilon_2 - 1)}. \tag{25}$$

In the limit that $(A_1/A_2) \to 0$, for instance, for a small convex object inside a very large enclosure, this reduces to the simple solution shown in Eq. (26).

$$\frac{q_{1-2}}{A_1} = \sigma\varepsilon_1\left(T_1^4 - T_2^4\right). \tag{26}$$

These are only two simple examples of the power of the radiation network approach to solving radiative heat transfer problems with many mutually irradiating surfaces.

The study of thermal radiative heat transfer goes on to consider radiative exchange between a gas and a heat transfer surface, complex radiation networks in absorbing and transmitting media, solar radiation and radiation within planetary atmospheres, and complex considerations of combined conduction–convection–radiation heat transfer problems. The reader is encouraged to investigate these and other topics in radiative heat transfer further.

IV. BOILING HEAT TRANSFER

The phenomenon of heat transfer from a surface to a liquid with a phase change to the vapor phase by the formation of bubbles is called boiling heat transfer. When a pool of liquid at its saturation temperature is heated by an adjacent surface which is at a temperature just above the liquid saturation temperature, heat transfer may proceed without phase change by single-phase buoyancy or natural convection.

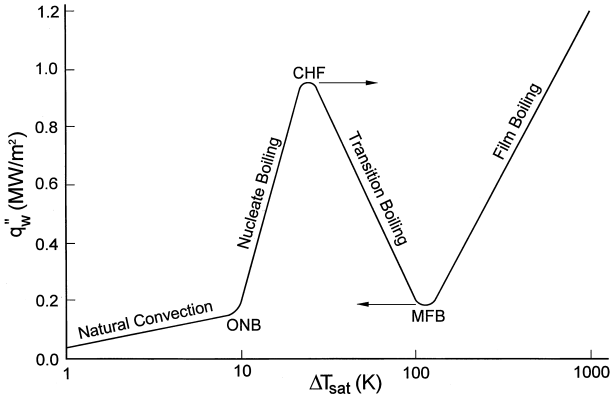


FIGURE 4 Pool boiling curve for water.

A. Onset of Pool Boiling

As the surface temperature is increased, bubbles appear on the heater surface signaling the onset of nucleation and incipient pool boiling. The rate of heat transfer by pool boiling as this is called is usually represented graphically by presenting the surface heat flux, q''_w, versus surface superheat, $T_w - T_{sat}$. This is referred to as the boiling curve. The process of boiling heat transfer is quite nonlinear, the result of the appearance of a number of regimes of boiling which depend fundamentally upon different heat transfer processes.

The components of the pool boiling curve have been well established and are shown graphically in Fig. 4. The first regime of the boiling curve is the natural convection regime, essentially a regime just preceding boiling, in which the heat transfer is by single-phase flow without vapor generation. In this regime, buoyancy of the hot liquid adjacent to the surface of the heater forces liquid to rise in the cooler liquid pool followed by fresh cold liquid passing over the heater to repeat the process.

B. Nucleate Boiling

A further increase in the surface superheat or the surface heat flux will drive the system to the onset of nucleate boiling (ONB), the point on the boiling curve at which bubbles first appear on the heater surface. The rate of vapor bubble growth, the area density of nucleation sites which become active, the bubble frequency and bubble departure diameter manifest themselves as dominant parameters controlling the heat flux as the pool enters the nucleate boiling regime, all of which are increasing functions of the surface superheat. Without further discussion, it is mentioned that the rate of heat transfer in the nucleate boiling regime is extremely sensitive to various properties and conditions including system pressure, liquid agitation, and subcooling; surface finish, age, and

coatings; dissolved noncondensible gases in the liquid; size and orientation of the heater; and nonwetting and treated surfaces. Heat fluxes in the nucleate pool boiling regime increase very rapidly with small increases in the surface superheat. The literature contains numerous efforts by various people to develop generalized correlations for nucleate pool boiling applicable to a wide range of liquids and generalized to include many of the properties and conditions previously listed. As a minimum, any successful correlation must include provisions which reflect the character or conditions of the heater surface as well as the properties of the boiling fluids themselves, requirements which have presented formidable obstacles to the development of any universally applicable correlation. One of the earliest attempts at such a correlation was developed by Rohsenow (1952), as seen in the following equation, for its historical significance and continued applicability:

$$\frac{c_\ell(T_w - T_{sat})}{i_{\ell g}} = C_{sf}\left[\frac{q''}{\mu_\ell i_{\ell g}}\left(\frac{\sigma}{g(\rho_\ell - \rho_g)}\right)^{0.5}\right]^{0.33} \mathrm{Pr}_\ell^s. \tag{27}$$

where $C_{sf} \sim 0.013$, $s = 1$ for water and $s = 1.7$ for all other fluids. Examination of this correlation reveals a theme underlying all of heat transfer and that is the essential requirement for accurate values of the physical and transport properties of the fluids of interest.

C. Critical Heat Flux

As the surface superheat in nucleate pool boiling continues to increase, the resulting increase in the boiling heat flux is accompanied by an increase in active nucleation sites on the surface of the heater, thus resulting in an increasing vapor production rate per unit area. The boiling heat flux will continue to increase up to a point at which the liquid can no longer remove any more heat from the surface due to vapor blanketing of the surface, restriction of liquid flow to the surface, and flooding effects which push liquid droplets away from the surface. There is no general agreement as to which of these mechanisms is responsible for the boiling crisis which ensues, and indeed each may be controlling under different geometric conditions. Regardless, soon the pool boiling curve reaches a peak heat flux which is called the critical heat flux (CHF). The critical heat flux in pool boiling is predominantly a hydrodynamic phenomenon, in which insufficient liquid is able to reach the heater surface due to the rate at which vapor is leaving the surface. As such, it is an unstable condition in pool boiling which should be avoided in engineered systems through design. There are two routes by which CHF can be reached. The first is by controlling the temperature of the heater surface, in which case the system will simply return to nucleate boiling if the superheat is reduced or enter into transition boiling if the superheat exceeds CHF.

D. Film Boiling

It is more likely, however, that in most engineering systems the actual independent variable would be the heat flux, not the surface temperature. In this case, any increase in the surface heat flux above the CHF limit would induce a huge temperature excursion in the surface as it became vapor-blanketed and heated-up adiabatically. This temperature excursion would continue until the imposed heat load was able to be transferred to the boiling liquid by thermal radiation from the surface almost exclusively. As a result of the vapor blanketing of the heater preventing liquid–solid contact, this boiling regime is called film boiling, and the occurrence of the thermal excursion from CHF into film boiling is known as burnout. This term comes from the fact that the resulting surface temperatures are, in general, so high that the surface and thus the equipment is damaged. Film boiling as a heat transfer process does not enjoy wide commercial application because such high temperatures are generally undesirable. In film boiling, a continuous vapor film blankets the heater surface which prevents direct contact of liquid with the surface. Vapor is generated at the interface between the vapor film and the overlying liquid pool by conduction through the vapor film and thermal radiation across the vapor film from the hot surface. It is of interest to note that in film boiling, the boiling heat flux is insensitive to the surface conditions unlike nucleate boiling, in which surface conditions or surface finish may play a dominant role. In transition boiling, the unstable regime between nucleate and film boiling, surface conditions do influence the data providing evidence that there is some liquid–surface contact in transition boiling which is not manifested in the film boiling regime. However, due to this decoupling of the boiling process from the heater surface conditions, film boiling is more tractable to analysis. The classical analysis of film boiling from a horizontal surface was performed by Berenson (1961). Many others have since contributed to the understanding of film boiling, notably by extending his work to very high superheats as well as to include liquid subcooling effects. The original model derived by Berenson for the film boiling heat transfer coefficient is reproduced in the following:

$$h = 0.425\left[\frac{k_g^3\rho_g(\rho_\ell - \rho_g)g(i_{\ell g} + 0.4c_{p,g}\Delta T_{sat}}{\mu_\ell \Delta T_{sat}(\sigma/g(\rho_\ell - \rho_g)^{1/2}}\right]^{0.25}. \tag{28}$$

For extension to higher temperatures at which thermal radiation becomes significant, a simple correction is made to the calculated film boiling heat flux by adding a radiative heat transfer contribution.

V. PHYSICAL AND TRANSPORT PROPERTIES

Accurate and reliable thermophysical property data play a significant role in all heat transfer applications. Whether designing a laboratory experiment, analyzing a theoretical problem, or constructing a large-scale heat transfer facility, it is crucial to the success of the project that the physical properties that go into the solution are accurate, lest the project be a failure with adverse financial consequences as well as environmental and safety implications. In the solutions of heat transfer problems, numerous physical and transport properties enter into consideration, all of which are functions of system parameters such as temperature and pressure. These properties also vary significantly from material to material when intuition suggests otherwise, such as between alloys of a similar base metal. Physical and transport properties of matter are surprisingly difficult to measure accurately, although the literature abounds with measurements which are presented with great precision and which frequently disagree with other measurements of the same property by other investigators by a wide margin. Although this can sometimes be the result of variations in the materials or the system parameters, all too often it is the result of flawed experimental techniques. Measurements of physical properties should be left to specialists whenever possible. It should come as no surprise, therefore, that the dominant sources of uncertainties or errors in analytical and experimental heat transfer frequently come from uncertainties or errors in the thermophysical properties themselves. This concluding section presents four tables of measured physical properties for selected materials under various conditions to illustrate the variability which can be encountered between materials and, in one case, the variability of a single material property as a function of temperature alone.

Table I presents the most frequently used physical and transport properties of selected pure metals and several common alloys at 300 K. Listed are commonly quoted values for the density, specific heat, thermal conductivity, and thermal diffusivity for 18 metals and alloys. Heat transfer applications frequently require these properties at ambient temperature due to their use as structural materials. For properties at other temperatures, the reader is referred to Touloukian's 13-volume series on the thermophysical properties of matter. It can be seen in Table I that some of the properties vary quite significantly from metal to metal. A judicious choice of metal or alloy for a particular application usually involves optimization of not only the thermophysical properties of that metal but also the mechanical properties and corrosion resistance.

Table II presents the most frequently used physical and transport properties for selected gases at 300 K. Once again, 300 K represents a temperature routinely

TABLE I Physical Properties of Pure Metals and Selected Alloys at 300 K

Material	T_{melt} (K)	ρ (kg/m^3)	c_p (J/kg · K)	k (W/m · K)	$\alpha \cdot 10^6$ (m^2/s)
Aluminum	933	2702	903	237	97.1
Bismuth	545	9780	122	7.9	6.6
Copper	1358	8933	385	401	117
Gold	1336	19300	129	317	127
Iron	1810	7870	447	80	23
304 SS	1670	7900	477	14.9	3.95
316 SS	~1670	8238	468	13.4	3.5
Lead	601	11340	129	35	24
Nickel	1728	8900	444	91	23
Inconel 600	1700	8415	444	14.9	4.0
Inconel 625	—	8442	410	9.8	2.8
Inconel 718	1609	8193	436	11.2	3.1
Platinum	2045	21450	133	72	25
Silver	1235	10500	235	429	174
Tin	505	7310	227	67	40
Titanium	1953	4500	522	22	9.3
Tungsten	3660	19300	132	174	68
Zirconium	2125	6570	278	23	12.4

TABLE II Properties of Selected Gases at Atmospheric Pressure and 300 K

Gas	ρ (kg/m^3)	c_p (kJ/kg · K)	$\mu \cdot 10^5$ (kg/m · s)	$\nu \cdot 10^6$ (m^2/s)	k (W/m · K)	$\alpha \cdot 10^6$ (m^2/s)	Pr
Air	1.18	1.01	1.98	16.8	0.026	0.22	0.708
Hydrogen	0.082	14.3	0.90	109.5	0.182	1.55	0.706
Oxygen	1.30	0.92	2.06	15.8	0.027	0.22	0.709
Nitrogen	1.14	1.04	1.78	15.6	0.026	0.22	0.713
CO_2	1.80	0.87	1.50	8.3	0.017	0.11	0.770

encountered in practical applications. The table considers five common gases and lists seven properties of general interest and frequent use. The reader is cautioned against the use of these properties at temperatures other than 300 K. All these properties with the exception of the Prandtl number are strong functions of temperature and significant errors can result if they are extrapolated to other conditions. The reader is once again directed to Touloukian for detailed property data.

Applications of heat transfer at very low temperatures such as at liquid nitrogen (76 K) and liquid helium (4 K) temperatures present unique challenges to the experimentalist and require knowledge of the cryogenic properties of matter. Table III presents a summary of the temperature-dependent specific heat of six common cryogenic materials over the temperature range from 2 to 40 K to illustrate the extreme sensitivity of this property in particular (and most cryogenic properties in general) to even slight variations in temperature. Clearly, experiments, analyses, or designs which do not use precise, accurate, and reliable data for the physical properties of materials at cryogenic temperatures will suffer from large uncertainties. In addition, precise temperature control is a necessity at these temperatures.

For research applications, these uncertainties could easily render experimental results and research conclusions invalid. The reader is cautioned to seek out the most reliable data for thermophysical properties when operating under cryogenic conditions.

Finally, there are occasions in heat transfer when it is advantageous to utilize liquid metals as a heat transfer medium. Table IV lists 12 low melting point metals commonly encountered in practice and lists their melting temperatures and boiling temperatures for comparison. The choice of a suitable liquid metal for a particular application does not provide for the flexibility which engineers and scientists have come to expect from other fluids at ordinary temperatures. Often the choice of an appropriate liquid metal for a particular application depends on the phase change temperatures as shown in Table IV. When and if no suitable liquid metal is found among the pure metals, alloys can be used instead. These alloys or mixtures usually have physical properties and melting/boiling temperatures which are significantly different from their constituent elements. No data for liquid metal alloys are given here; indeed, the data for liquid metal alloys are meager.

It is crucial to the success of any heat transfer experiment, analysis, or facility that careful and judicious choices are made in the selection of the materials to be

TABLE III Specific Heat of Selected Materials at Cryogenic Temperatures

	Specific heat (J/kg · K)					
T (K)	Al	Cu	α-Iron	Ti	Ice	Quartz
2	0.05	0.0066	0.183	0.146	0.12	—
4	0.26	0.0217	0.382	0.317	0.98	—
6	0.50	0.0545	0.615	0.540	3.3	—
8	0.88	0.114	0.900	0.840	7.8	—
10	1.4	0.205	1.24	1.26	15	0.7
15	4.0	0.663	2.49	3.30	54	4.0
20	8.9	1.76	4.50	7.00	114	11.3
30	31.5	6.53	12.4	24.5	229	22.1
40	77.5	14.2	29.0	57.1	340	65.3

TABLE IV Melting and Boiling Temperatures of Some Common Liquid Metals

Metal	T_{melt} (K)	T_{boil} (K)
Lithium	452	1590
Sodium	371	1151
Phosphorus	317	553
Potassium	337	1035
Gallium	303	2573
Rubidium	312	969
Indium	430	2373
Tin	505	2548
Cesium	302	1033
Mercury	234	630
Lead	601	2023
Bismuth	544	1833

used and that the data on their thermophysical properties are accurate and precise. It is difficult and expensive to determine these properties on an application by application basis, and property data which have not been measured by specialists may suffer large uncertainties and errors. The results of experiments and analyses can only be as accurate and reliable as their data and frequently that accuracy and reliability are limited by the thermophysical properties which were used.

SEE ALSO THE FOLLOWING ARTICLES

CRYOGENICS • DIELECTRIC GASES • ELECTROMAGNETICS • FUELS • HEAT EXCHANGERS • HEAT FLOW • THERMAL ANALYSIS • THERMODYNAMICS • THERMOMETRY

BIBLIOGRAPHY

Berenson, P. J. (1961). "Film boiling heat transfer from a horizontal surface," *J. Heat Transfer* **83,** 351–358.

Eckert, E. R. G., and Drake, R. M. (1972). "Analysis of Heat and Mass Transfer," McGraw-Hill, New York.

Hartnett, J. P., Irvine, T. F., Jr., Cho, Y. I., and Greene, G. A. (1964–present). "Advances in Heat Transfer," Academic Press, Boston.

Rohsenow, W. M., Hartnett, J. P., and Ganic, E. N. (1985). "Handbook of Heat Transfer Fundamentals," McGraw-Hill, New York.

Rohsenow, W. M. (1952). "A method of correlating heat transfer data for surface boiling of liquids," *Trans ASME* **74,** 969.

Sparrow, E. M., and Cess, R. D. (1970). "Radiation Heat Transfer," Brooks/Cole Publishing Company, Belmont, CA.

Touloukian, Y. S., *et al.* (1970). "Thermophysical Properties of Matter," TPRC Series (v. 1–13), Plenum Press, New York.

VanSant, J. R. (1980). "Conduction Heat Transfer Solutions," Lawrence Livermore National Laboratory, UCRL-52863.



Liquids, Structure and Dynamics

Thomas Dorfmüller
University of Bielefeld

I. Introduction
II. Structure of Liquids
III. Dynamic Properties of Liquids
IV. Molecular Interactions and Complex Liquids
V. Compartmented Liquids
VI. Glass-Forming Liquids
VII. Gels
VIII. Dynamics in Complex Liquids
IX. Conclusions

GLOSSARY

Amphiphiles Molecules consisting of a hydrophobic and a hydrophilic moiety.

Complex liquids Liquids that consist of molecules whose anisotropic shape, specific interactions, and intramolecular conformations determine their properties to a significant degree.

Dynamic spectroscopy Spectroscopic technique that uses the shape of spectral lines to obtain information about the dynamics of molecules.

Gels Polymer–liquid mixtures displaying no steady-state flow. Gels are cross-linked solutions.

Glass Solidlike amorphous state of matter.

Liquid crystals Liquids consisting of highly anisotropic molecules whose orientations are strongly correlated.

Micelles Aggregates formed in an aqueous solution of a detergent.

Microemulsions Ternary solution of water, oil, and a detergent that forms dropletlike aggregates.

Pair correlation function Function of distance r that describes the probability of finding a molecule at a place located at a distance r from the origin, given that another molecule is located at the origin. See Fig. 1.

Phase diagram Diagram having the coordinates pressure and temperature, in which the solid, liquid, and gaseous phases occupy different regions.

Plastic crystals Crystalline phases displaying a liquidlike rotational mobility of the molecules.

Simple liquids Liquids consisting of spherical or nearly spherical molecules interacting with central forces and without conformational degrees of freedom.

Time correlation function Statistical quantity used to describe the temporal evolution of a random process. See Eq. (13).

THE LIQUID STATE is a condensed state of matter that is roughly characterized by high molecular mobility and

a low degree of order compared with solids. Liquids can be distinguished from gases by their high density.

Liquid structure is studied by various scattering methods and liquid dynamics by a large number of spectroscopic techniques.

The liquid phase of matter is of paramount importance in physics, chemistry, biology, and engineering sciences. Especially, liquid or liquidlike systems such as micelles, microemulsions, gels, and membranes play important roles in biology and in industrial applications. Although much detailed knowledge of liquids has been accumulated, the study of many fundamental issues in liquid-state physics and chemistry is actively pursued in many fields of science.

I. INTRODUCTION

The liquid state of matter cannot be easily defined in an unambiguous and consistent way. It is often defined in terms of the phase-diagram (i.e., with respect to the solid and gaseous state). However, the distinction between the liquid phase and the gas phase is not sharp in the critical region of the phase-diagram, and the distinction of a liquid from a solid is also unclear for substances that have a tendency to supercool and are able to form glasses at low temperatures. Furthermore, many fluid substances are known that display specific structures, such as liquid crystals and micelles, where some of the criteria usually attributed to liquids do not apply. One could give a wide, but still to some extent ambiguous, definition of liquids by saying that a liquid is a disordered condensed phase. This would then include glasses, which due to the low mobility of the constituent molecules are usually regarded as amorphous solid systems. Another case where the limits of the liquid state are ill-defined is that of disordered clusters of molecules and of two-dimensional disordered arrangements on surfaces. The question of whether these phases should be considered liquids is a matter of the context in which they are studied.

Liquids can be classified according to the properties of the molecules that constitute them. We thus distinguish between atomic and molecular liquids, among nonpolar, polar, and ionic liquids, and between liquids whose molecules do or do not display hydrogen bonding. Since interparticle interactions play a central role in determining the properties of liquids, we can broadly classify simple and complex liquids according to the way in which the molecules or atoms of the liquid interact. The simplest liquids are those consisting of atoms of noble gases. Thus, liquid argon, being considered as the prototype of a simple liquid, has been the object of many studies because of the absence of any complicating features in the intermolecular interaction. What distinguishes, for example, liquid argon from most other liquids is the spherical shape of its atoms leading to central interaction forces, the dispersive character of the interparticle forces, and the absence of internal degrees of freedom.

Other liquids that can be considered simple should consist of atoms or molecules with shapes not deviating much from a sphere; they should not display noncentral, angle-dependent, or specific saturable interparticle forces like, for example, those that lead to the formation of hydrogen bonds, and, finally, the internal degrees of freedom, especially configurational degrees of freedom, should not much influence the properties of the liquid. According to these criteria, the majority of liquids are complex, the concept of a simple liquid being the result of an extrapolation of the properties of a relatively small number of liquids whose molecules comply to some extent to the above-mentioned requirements of simple liquids. The concept of a simple liquid has been very fruitful in contributing to the development of the concepts that are necessary to describe the essentials of the liquid state. However, since most interesting and important liquids must be reckoned among the complex liquids, the study of these is extremely important.

II. STRUCTURE OF LIQUIDS

The ordered structure of crystalline solids is a consequence of interparticle interactions leading to a dependence of the free energy of a system of interacting particles on their arrangement in space. The stability of a particular crystalline structure at thermodynamic equilibrium results from the minimum of the free energy achieved in this state. Because such interactions are negligible in gases at low pressure, we observe chaos instead of order in gaseous systems. The case of liquids is intermediate between the two extremes of perfect order in ideal crystalline solids and complete lack of order in ideal gases. More precisely, the molecules of the liquids interact, and as a consequence they tend to arrange themselves in a constrained structure. On the other hand, the thermal energy of a liquid is high enough so that the molecules rearrange themselves rapidly and continuously. As a result, order in most liquids is not constant in time, nor does it extend over distances larger than a few molecular diameters. If the shapes of the molecules and the attractive forces between molecules are anisotropic, then we may observe orientational order in a liquid, with the molecules with aligning themselves so that the orientation of neighboring molecules is not completely random. The opposite case is observed in plastic crystals, which are positionally ordered systems in which the molecules are, however, relatively free to rotate; so we do not have orientational order as in classical crystals.

Both translationally and rotationally ordered structures can be described by appropriate functions of spatial variables such as interparticle distance and relative orientation. Translational order is described by the radial pair correlation function $g(r)$ and orientational order by the static orientational correlation function. The function $g(r)$ is the probability of finding a molecule at a point A at a distance located between r and $r + dr$ from another point B, given that another molecule is at B. The static orientational correlation is a number reflecting the probability that a molecule located at a distance r from another molecule is oriented in such a way that the two molecules form an angle between ϑ and $\vartheta + d\vartheta$. For this definition to apply, the molecules must have a symmetry allowing the definition of a physically identifiable orientation axis. This is the case, for example, with linear and cylindrically symmetric molecules. Thus, the size of this quantity gives us a clue as to whether molecules located at a distance r from each other tend to align in parallel, antiparallel, or perpendicular orientations and characterizes the average angle between such molecules.

The importance of both quantities stems from the fact that they can be measured by the diffraction of electromagnetic radiation and by scattering of slow neutrons, both having a wave-length of approximately the average interparticle distance. Figure 1 displays a characteristic shape of $g(r)$, illustrating the exclusion of neighbors at small distances from the reference molecule. The first, more pronounced, peak at r_1 corresponds to the shell of the nearest neighbors. The second at r_2 and the further peaks come from more distant and hence more diffusely distributed shells. The oscillations characteristic of the radial distribution decay after a small number of maxima and minima showing that no long-range order is present at distances

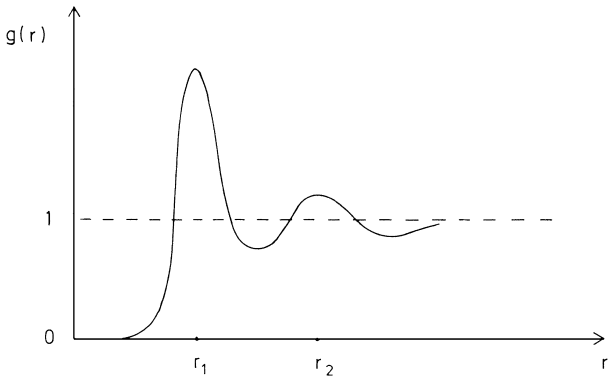


FIGURE 1 A typical radial pair correlation function of a simple liquid. Note the maxima at r_1 and r_2, which illustrate the increased probability of finding a molecule at these distances from the central molecule. Note also the value 0 at small distances and the limiting value of 1 at large distances. The first is a consequence of the repulsive interaction of molecules, and the second illustrates the randomization of the mean particle density at large distances.

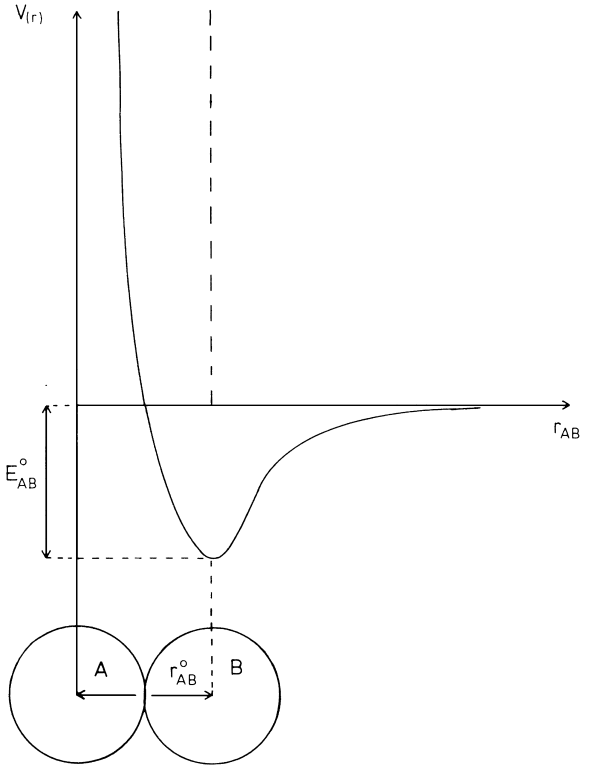


FIGURE 2 The pair potential energy between two molecules A and B. The two molecules, represented by the spheres, are shown at the equilibrium distance r_{AB} if the potential is a Lennard–Jones potential. In terms of the hard core repulsive potential (vertical dashed line) this is the contact position with the center-to-center distance equal to r_{AB}. The Lennard–Jones potential: $V(r) = 4E_{AB}[(r^0/r)^{12} - (r^0/r)^6]$ results from the superposition of the r^{-12} repulsive branch (upper part of the solid curve) and the r^{-6} attractive branch (lower half of the solid curve).

as large as a few molecular diameters. For large values of the distance r the product $\rho g(r)$ approaches the value of the average number density $\rho = N/V$ of the equilibrium distribution, where N is the number of molecules contained in the volume V. In contrast to this, a system with long-range order would display nondecaying oscillations of $g(r)$ over significant distances.

Although the general form of the intermolecular forces is known, it is very difficult to derive directly from this approximate knowledge the exact shape of the radial distribution function. As a useful approximation, however, the repulsive branch of the potential has been approximated by a hard core potential and the attractive branch expressed by simple inverse power of the intermolecular center-to-center distance. This is illustrated in Fig. 2. With such an approximation it was found that the main features of the radial pair distribution function can be explained qualitatively even if we completely neglect attraction. It thus appears that most of the liquid structure is the result of the steep repulsive intermolecular pair potential.

Steep in this context means a potential energy function that can be expressed as an inverse power of the intermolecular distance with an exponent that is significantly larger than the value $n = 6$ found in dispersion forces. Usually, steep potentials are approximated by an exponent $n = 12$ or larger because this is the case for the repulsive part of the Lennard–Jones potential illustrated in Fig. 2. In most cases the repulsive branch of the potentials was shown to be much steeper than the attractive branch, although some complex liquids, such as associating liquids, do have steep attractive branches that are essential in determining their structure. In this latter case the potential has the shape of a narrow, steep-walled well leading to relatively stable dimers or higher aggregates.

In a system of particles interacting through central pair forces, we can derive a simple equation between the excess internal energy U per molecule due to intermolecular interactions, the pair potential $V(r)$ and the radial distribution function $g(r)$:

$$U = N2\pi\rho \int_0^\infty g(r)V(r)r^2\,dr. \tag{1}$$

We can also derive an expression for the equation of state for this system in terms of the radial distribution function and the gradient of $V(r)$:

$$PV = NkT\left[1 - \frac{1}{6kT}\int_V g(r)r\frac{dV(r)}{dr}\,dr\right]. \tag{2}$$

In this equation k is the Boltzmann constant and P the pressure. The compressibility equation, Eq. (3), connects the isothermal compressibility defined as $\beta_T = (\partial V/\partial P)_T V^{-1}$ with the radial pair distribution function:

$$kT\beta_T = 1 + \rho\int_V [g(r) - 1]\,dr. \tag{3}$$

The liquid structure that is inherent in $g(r)$ can also be described in terms of the structure factor $S(\mathbf{k})$, which is a quantity used to describe neutron and X-ray scattering experiments. The elastic scattering of, for example, neutrons having a typical wavelength of 1 Å is determined by the local arrangement of the scattering atoms. The structure factor is connected to the radial distribution function by means of the equation

$$S(\mathbf{k}) = 1 + \rho\int_V \exp(-i\mathbf{kr})[g(r) - 1]\,d\mathbf{r}. \tag{4}$$

This equation shows that the structure factor can be expressed in terms of the Fourier transform in space of the radial distribution function. For isotropic liquids the radial pair distribution function is a function of the modulus r and the structure factor of the modulus of the vector $\mathbf{k}$. If the liquid is anisotropic, we must use instead the full vectors $\mathbf{r}$ and $\mathbf{k}$.

Actually, $S(\mathbf{k})$ describes the liquid structure in k-space, which is the reciprocal of ordinary r-space. The role of this function in describing the structure of the liquid is determined by the character of the incident radiation which is characterized by its wavevector $\mathbf{k}_i$ and that of the scattered radiation by $\mathbf{k}_s$. The conservation of momentum of the system, liquid + incident radiation + scattered radiation, leads to a scattering intensity for a given angle of observation that depends only on $S(\mathbf{k})$, where the vector $\mathbf{k}$ is defined by

$$\mathbf{k} = \mathbf{k}_s - \mathbf{k}_i. \tag{5}$$

$S(\mathbf{k})$ curves can be calculated by model theories that can then be tested against the experimental curves obtained from neutron scattering experiments.

In the cases of linear (e.g., nitrogen and carbon disulfide) and tetrahedal molecules (e.g., yellow phosphorus and carbon trichloride), the diffraction methods have shown that the former indeed tend to align in the liquid, whereas the latter in some cases form interlocked structures. The structural information obtained by diffraction methods is important but far from complete, and the confirmation by model calculations is essential. The calculation of accurate structure factors from diffraction experiments is often hampered by correction problems and problems of interpretation. The use of isotopically substituted molecules has proved essential in obtaining the necessary data to calculate the more detailed atom–atom pair distribution functions of molecular liquids.

With increasing complexity of the molecules, the problems increase too. In the case of some more complex molecules, however, such as acetonitrile, chloroform, methylene chloride, and methanol, the diffraction methods have given structure factors that compare favorably with theoretical data. Especially, one can confirm the formation of hydrogen-bonded chainlike structures, which are expected from the physical properties of these substances and from some dynamic data. The above relations can be extended to describe more complex polyatomic molecular liquids if appropriate parameters for the description of the molecular coordinates (i.e., either the relative position of the center of mass r and the angles describing the orientation of the molecule or the set of parameters specifying the position of all the atoms in the molecule) are introduced. The radial pair distribution function then becomes a function of all these coordinates and is generally much too complex to calculate or even to visualize.

A very useful simplification of the description and hence the calculation of liquid structure using site correlation functions was introduced with the so-called RISM theory. This theory incorporates the chemical structure of the molecules into the model by approximating them to objects consisting of hard fused spheres modeling their

chemical structures. The usefulness of this view in describing the pair correlation function stems from the fact that it is mainly the shape of the molecule that is critical in determining the structure of the liquid.

If, on the other hand, the role of the long-range attractive part of the intermolecular potential is important (e.g., in the calculation of thermodynamic properties or in the case of strongly structured liquids such as water or some other polar liquids), different methods must be used, such as perturbation theories. These are based upon the assumption that we can split the intermolecular interactions into a simple well-defined part and into another that is considered a small perturbation, but which confers to the more complex system its specific properties, which we want to calculate.

The main problem lies in the question whether it is possible to find an adequate reference-state leading to a known structure so that the interactions of the more complex liquid can be obtained by adding a perturbation term to the reference potential. The hard-sphere potential was often chosen as a reference, but its usefulness for polyatomic liquids has been seriously questioned.

The strengths and limitations of theoretical models that are used to obtain a quantitative description of liquid structure are often assessed by comparing the results with X-ray and neutron diffraction data and with the results of computer simulation calculations. The latter provide us with a method of calculating numerically the properties of model liquids consisting of molecules with a well-defined intermolecular potential. In molecular dynamics computer simulations, the classical trajectories of an ensemble of molecules are calculated by solving the equation of motion. The static properties (i.e., pair distribution functions, equations of state, and internal energy) are calculated as averages over a sufficiently large number of equilibrium configurations created from the trajectories. The value of computer simulation lies in the possibility of calculating separately the effects of different features of real molecules: shape, the potential energy parameters, the dipole moment, and several others. This proved to be very valuable for the understanding of the important factors affecting the structure of liquids. On the other hand, our incomplete knowledge of the intermolecular potential of real molecules prevents the computer simulations from giving an exact replica of the real liquid. Furthermore, due to the necessary restrictions in computer capacity, the ensembles that can be reasonably handled consist of 1000 molecules or less. The averages over such ensembles are considered to represent, to a sufficient degree of accuracy, statistical averages in a bulk liquid consisting of some 10^{20} molecules. This entails problems of a statistical nature that have been only partially solved.

It is fair to say that our knowledge of the structure of liquids has advanced in the last two decades, the essential mechanisms determining liquid structure being understood in principle. What is still lacking is an accurate knowledge of intermolecular potentials derived either from experiments in the liquid state or by ab initio quantum-mechanical calculations. Furthermore, we must be aware that most of our models are approximations and that we still are unable to predict whether a given approximation is adequate to give good thermodynamic, structural, or dynamical data. We also do not know why most of the currently used models give good results for data of one of the above-mentioned classes but poor results for another class.

III. DYNAMIC PROPERTIES OF LIQUIDS

The description of the equilibrium state of a liquid by means of the radial pair distribution function or the structure factor can be extended to include time-dependent properties of liquids. This can be done by the use of the Van Hoove correlation function $G(r, t)$ which has been introduced as a tool for the description of quasi-elastic neutron-scattering results. This function is both space- and time-dependent. In analogy to the definition of the static radial distribution function by means of Eq. (1), the Van Hoove correlation function is defined as a time-dependent density–density correlation function:

$$G(r, t) = \frac{\langle \rho(r, t) \cdot \rho(0, 0) \rangle}{\langle \rho(0, 0) \rangle^2}, \tag{6}$$

where $G(r, t)$ is the probability of finding a particle i in a region dr around a point r at time t, given that there was a particle j at the origin at time $t = 0$.

To separate the motion of particles in a laboratory-fixed frame of reference from the relative motion of the particles, it is convenient to separate $G(r, t)$ into a self and a distinct part:

$$G(r, t) = G_s(r, t) + G_d(r, t). \tag{7}$$

Figure 3 illustrates the behavior of $G_s(r, t)$ and $G_d(r, t)$ on three time scales. The time scales are considered with respect to the so-called structural relaxation time τ, which is defined as the average time required to change the local configuration of the liquid. In Fig. 3, the following cases are distinguished:

1. The time scale is short with respect to the structural relaxation time.
2. The time scale is similar to the structural relaxation time.

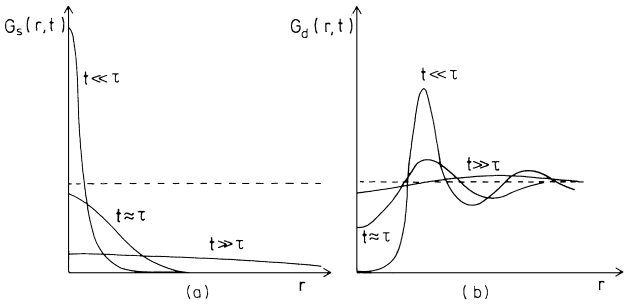


FIGURE 3 The shapes of (a) the self term and (b) the distinct term of the Van Hoove space-time correlation function. At times that are short with respect to the structural relaxation time, $G_s(r, t)$ is sharply peaked since the reference molecule did not have time to change its position significantly, whereas $G_d(r, t)$ is zero due to the repulsion of the reference molecule at the origin, which does not allow another molecule to occupy the same position. With increasing time, as t becomes similar to τ, $G_s(r, t)$ broadens because the probability of finding the reference molecule away from the origin increases. On the other hand, at small distances from the origin, $G_d(r, t)$ increases as the probability of finding another molecule at the origin is no longer zero. Finally, at long times ($t \gg \tau$), the probability of finding a given molecule at a distance r is small and independent of the distance from the origin, and the probability of finding some molecule at r is 1.

3. The time scale is long with respect to the structural relaxation time.

One sees that at short times G_s is sharply peaked about $r = 0$, whereas G_d displays the characteristic oscillations similar to the time-in-dependent radial pair distribution function in Fig. 1. At long times, both functions vary little in space and approach the steady-state value as the local distribution is nearly averaged out at times long as compared to τ.

The described generalization leading to the space- and time-dependent Van Hoove correlation function is readily extended to the structure factor which then becomes frequency dependent. The use of a frequency dependent dynamic structure factor $S(k, \omega)$ stems from the study of the spectra of scattered slow neutrons. The symbol ω is for the angular frequency. Thermal neutrons are well suited to the study of the dynamics of liquids because their energy is comparable to kT, and the wavelength associated with the neutrons is comparable to intermolecular distances at liquid densities. The measurable derivative $d^2\sigma/dO\,d\vartheta$ of the differential cross section $d\sigma/dO$ is directly related to $S(k, \omega)$ by the equation

$$\frac{d^2\sigma}{dO\,d\vartheta} = b^2\left(\frac{k_1}{k_0}\right)S(k, \omega), \tag{8}$$

where σ is the total cross section and b the scattering length typical for the scattering atom (of the order of magnitude of the nuclear radius). If the molecule is heteronuclear, this relation has the form of a sum over all j atomic species with scattering lengths b_j, O represents the solid angle under which the scattered radiation is detected and k_1 and k_0 the moduli of the wavevectors of the neutrons before and after the scattering event, respectively.

The dynamic structure factor can be separated into a self and a distinct part $S_s(k, \omega)$ and $S_d(k, \omega)$ corresponding to the self and distinct parts of the Van Hoove correlation function. This separation acknowledges the fact that the molecular motion detected by neutron scattering involves both single-particle and collective motions. We can use two extreme models to describe the situation.

A. The Perfect Gas Model

The assumption of a free motion of the molecules with a mass M and the most probable velocity $v_0 = \sqrt{2kT/M}$ leads to the following expression for the Van Hoove correlation function:

$$G_s(r, t) = \left(\pi^{-2}v_0\tau\right)^{-3}\exp\frac{-r^2}{(v_0\tau)^2} \tag{9a}$$

$$G_d(r, t) = \sigma. \tag{9b}$$

From this expression we can obtain the following expression for the dynamic structure factor for this model:

$$S(k, \omega) = \left(kv_0\pi^{1/2}\right)^{-1}\exp\frac{-\omega^2}{(kv_0)^2}. \tag{10}$$

In liquids, this limit of a free (i.e., collisionless) motion is realized in the limit $r \to 0$ and $t \to 0$ corresponding to $k \to \infty$ and $\omega \to \infty$. Such behavior is approximated in scattering experiments with wavelengths significantly shorter than the average interparticle spacing of a few Angstroms (thermal neutrons) and probing times shorter that the average time between successive collisions.

B. Single-Particle Motion in the Hydrodynamic Limit

The opposite extreme to this limit is obtained at long times and large distances corresponding to $k \to 0$ and $\omega \to 0$. In this range the molecular interactions and not their masses determine the motion that is now monitored by the scattering of long-wavelength radiation (e.g., light scattering). This latter range of liquid dynamics corresponds to the hydrodynamic limit. In this limit the liquid can be treated as a continuum to which the hydrodynamic equations apply, the molecular details being formally introduced as extensions of the classical Navier–Stokes equations (e.g., by using a frequency-dependent viscosity to take care of molecular relaxation processes). In the hydrodynamic limit (i.e., when r and t are sufficiently large), the single-particle correlation function obeys a diffusion equation

similar to Fick's differential diffusion equation. Under appropriate boundary conditions, we obtain the following integrated form:

$$G_s(r, t) = (4\pi Dt)^{-3/2} \exp \frac{-r^2}{4Dt}. \tag{11}$$

This is also Gaussian, differing, however, in the time dependence from the case of the free motion in Eq. (10). The corresponding dynamic structure factor is given by

$$S_s(k, \omega) = \frac{(1/\pi)Dk^2}{\omega^2 + Dk^2}. \tag{12}$$

This expression represents the spectrum of the scattered intensity at a fixed value of the wavevector (i.e., at a fixed angle of observation).

The time correlation function formalism has been shown to be adequate for representing liquid dynamics in a convenient way. Thus, some experimental methods such as photon correlation spectroscopy directly give time correlation functions; others such as infrared and Raman bandshape analysis operate in the frequency domain, and the obtained spectra can be Fourier transformed to give time correlation functions. Figure 4 visualizes the relation between the two domains. This procedure is based on the fluctuation-dissipation theorem of statistical mechanics which connects random thermal fluctuations in a medium to the power spectrum characterizing the frequency spectrum of the process. A time correlation function of a dynamical molecular variable $A(t)$ (e.g., a dipole moment) is defined by

$$C(t) = \langle A(0)A(t) \rangle. \tag{13}$$

The correlation time τ of a process described by the above correlation function is defined as the integral of the correlation function over the time from $t = 0$ to $t = \infty$. If the process described by the above correlation function is diffusive, then the correlation function is exponential:

$$C(t) = C(0) \exp \frac{-t}{\tau}. \tag{14}$$

The time constant τ describing the decay of the correlation is termed a relaxation time. The corresponding spectrum $I(\omega)$ has a Lorentzian shape given by

$$I(\omega) = \frac{1}{1 + (\omega - \omega_0)^2 \tau^2}. \tag{15}$$

The Lorentzian bandshape as indicated in the figure is characterized by the relaxation time that thus can be extracted from the half width at half height Γ of the spectral band by

$$\tau = \frac{1}{2\pi\Gamma}. \tag{16}$$

The spectral band function $I(\omega)$ and the corresponding correlation function $C(t)$, as illustrated in Fig. 4, are a Fourier transform pair (i.e., they can be uniquely transformed into each other).

The macroscopic transport coefficients such as the mass diffusion coefficient, the thermal conductivity coefficient, and the macroscopic shear viscosity have been related to the time integral of pertinent correlation functions. Thus, the mass diffusion coefficient D is given by

$$D = \frac{1}{3} \int_0^\infty \langle \mathbf{v}(0)\mathbf{v}(t) \rangle \, dt. \tag{17}$$

In this equation $\mathbf{v}(t)$ is the molecular velocity at time t.

We have presently at our disposition several sources of information about the molecular dynamics of liquids. Among them the most important experimental techniques are Rayleigh and Raman light scattering; infrared and far infrared spectroscopy; NMR spectroscopy; fluorescence anisotropy methods, either stationary or time dependent; and time-dependent spectroscopy from the nanosecond to the femtosecond time scale. On the other hand, one of the most important sources of dynamical information on liquids is the computer simulation by means of molecular dynamics. The method aiming at extracting dynamical information from the shape of spectra is termed dynamical spectroscopy. The dynamical information contained in spectral band-shapes is in most cases complex, since the spectra reflect rotational, translational, and vibrational broadening mechanisms that cannot be uniquely sorted out. Each spectroscopic method has its strengths and its inherent restrictions, which is why most of the progress has been obtained by the simultaneous application of several methods on the same liquid. One of the most serious restrictions is that some methods (e.g., Rayleigh scattering) probe collective motions that are very difficult to relate

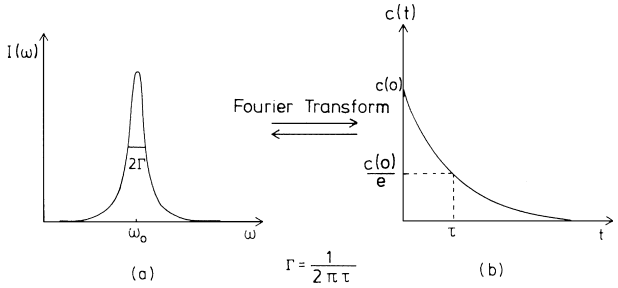


FIGURE 4 Relation between (a) the power spectrum (frequency domain) of a relaxation process and (b) the correlation function (time domain) of a dynamical variable describing this process. In dynamic spectroscopy, the half width at half height of the spectral line is measured, and the correlation function is obtained by Fourier transforming the spectral profile. The relaxation time can be directly calculated from the linewidth by the relation indicated in the figure if the spectral profile is a Lorentzian. In photon correlation spectroscopy, which operates in the time domain, a correlation function is directly measured.

exactly to single molecule motions. On the other hand, collective motions, essential to the understanding of liquid dynamics, are of great interest in themselves. On the other extreme, NMR data probe essentially single molecule motions, however, in contrast to optical spectroscopic methods, they do not give all the information contained in full time correlation functions but only correlation times.

In all the dynamical methods, we must be aware of instrumental restrictions with regard to the accessible time scale. The extension of the time scale over as many decades as possible is vital to the understanding of the underlying molecular mechanism and is one of the main experimental goals in this area. For this reason it is important to obtain reliable data with different methods on time scales complementing each other and which can be properly adjusted to each other. This is for example the case with the simultaneous study of the depolarized light scattering and flourescence anisotropy of the same label molecule dissolved in a liquid which can be used to monitor molecular rotations on a time scale extending from a few picoseconds to approximately 1 μs (i.e., over six decades).

In the last years the use of ultrashort laser pulses has become an increasingly important tool for the study of liquid dynamics. By this method it has become possible to study directly in the time-domain processes taking place in the picosecond and subpicosecond time range. These include orientational processes in liquids, the rates of charge transfer processes, the rate of recombination of ions to molecules in a liquid cage, and a number of solvent-dependent photophysical processes. Thus, vibration–rotation coupling and the rates of vibrational energy and phase relaxation were studied by picosecond spectroscopy, and the corresponding rates could be determined in some cases. Although the data obtained in the time domain and those in the frequency domain are rigorously linked, the former sometimes allow us to circumvent serious instrumental complications. In combination with spectroscopic lineshape analysis, real-time techniques have improved significantly our understanding of liquid-state dynamics.

Intermolecular dynamics is manifested in so-called interaction-induced spectra. This phenomenon, which leads to the occurrence of forbidden spectral lines appearing in high-density gases and in liquids has been studied extensively in the last years and has been shown to be helpful in obtaining information mainly about the short-time dynamics of liquids. The main mechanism by which these spectra are produced is the induction of a time-dependent dipole on a molecule by electric fields of other molecules in its immediate neighborhood and the interaction of this dipole with the electric field of the light.The intermolecular inducing fields may be coulomb-, dipole-, or higher-multipole fields, as the case may be. Furthermore, high-energy collisions produce distortions of the colliding molecules, thus inducing transient dipoles that also contribute to interaction-induced spectra. It appears that the wealth of information concealed in interaction-induced spectra is presently the main problem encountered in the analysis of such data.

IV. MOLECULAR INTERACTIONS AND COMPLEX LIQUIDS

The structures as well as the dynamics of liquids in equilibrium are determined by interactions of the molecules. Knowledge of these interactions is essential in obtaining a theoretically founded description of the physics of liquids. However, it is still very difficult to carry out quantum mechanical ab initio calculations of the intermolecular potential, although the basic understanding of the interactions between molecules is available. Thus, such calculations have been fruitful only for a small number of molecules consisting of a relatively small number of atoms.

The alternative to theoretical calculations is to determine intermolecular potentials by accurate gas-phase experiments that probe essentially two-particle interactions. However, this has also proved at least ambiguous since it is generally not possible to use unmodified gas-phase potentials for liquids. On the other hand, the reverse method of determining intermolecular potentials from liquid-state data is also unyielding because the method is model-dependent and, additionally, because most measurable quantities present themselves as integrals over ensembles of molecules from which the integrand cannot be uniquely determined.

For the present we must use empirical potentials that represent averages over those effects that cannot yet be explicitly taken into account. A major problem that is still unsolved is the calculation of many-particle interactions in an ensemble of interacting molecules. This is critical for the description of a liquid since at liquid densities, due to the small distances between interacting molecules, we are, in principle, not allowed to express the total interaction in the liquid as a sum of pairwise additive interactions, neglecting the many-body character of the problem. Some experimental results have been interpreted by assuming that many-particle interactions are important; however, such interpretations are still far from being quantitative.

V. COMPARTMENTED LIQUIDS

The nature of molecular interactions is very important in determining the structure and the properties of complex liquids. Many molecules consist of two different parts, the

one interacting strongly with water (the hydrophile) and the other not (the hydrophobe). In aqueous solutions we observe compact or lamellar liquid aggregated structures, depending on the nature of the solute, the temperature, and the concentration. The micelles thus formed are assumed to have a rather compact hydrophobic core surrounded by a hydrophilic shell. The hydrophiles are generally ionic or polar groups, whereas the hydrophobe is often an aliphatic chain. The driving force of aggregation in water of such molecules, called amphiphiles or detergents, is the minimization of the total free energy resulting from contributions from the water–water, water–hydrophile, and hydrophile–hydrophile interactions. The phase diagram of amphiphile–water mixtures displays several distinct phases with quite different properties, depending on the temperature and the concentration of its constituents. The structure of the resulting aggregates is a function of the nature of the hydrophile and of the length of the hydrophobic chain (see Fig. 5). Also the structure of the liquid within the micellar core seems to be in some cases different from that in bulk liquids with the effect that solubilized species may display a specific behavior as regards reactivity, acidity, mobility, etc.

Another example of compartmentation of practical importance is the case of microemulsions, which are formed when water, oil, and a detergent are mixed in appropriate proportions. Such systems are used to solubilize otherwise unsoluble substances and to promote chemical reactions by capturing them in their interior and thus increasing the local concentration of the reactants. Catalytic reactions in micelles and microemulsions play an increasing role in chemistry.

The occurrence of localized compartmented liquidlike phases is a very important phenomenon and plays a major role in biological systems where both fluidity and compartmentation are essential. The internal fluidity of compartmented liquid phases is studied intensively by spectroscopic methods such as ESR spectroscopy and fluorescence anisotropy decay of convenient dissolved or chemically bound labels.

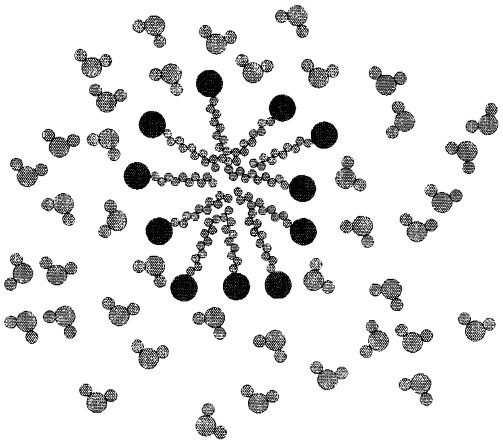


FIGURE 5 The structure of a micelle in water.

Even in the absence of distinct phases, molecules that consist of groups with different affinities to the solvent give rise to more localized and less randomized structures that affect the physical and chemical properties of liquid mixtures. Hydrogen-bonding molecules can belong to this category. Charge transfer interactions may affect in a similar fashion the local structure of a liquid mixture.

VI. GLASS-FORMING LIQUIDS

Amorphous substances with a solidlike rigidity play an important technological role, and their study is one of the major fields in materials science. Such substances are generally obtained when a liquid is cooled below its melting point while preventing crystallization. Glass-forming liquids (i.e., those that can be obtained in the glassy state) must have special properties connected with the symmetry, configuration, and flexibility of the molecules or their ability to form intermolecular bonds. Polymeric liquids are among the best studied glass-forming systems.

The properties of liquids at different temperatures can basically be understood in terms of the kinetic energy and the intermolecular potential of the molecules. In some cases, however, during the process of cooling, the viscosity of a liquid increases by several orders of magnitude in a rather narrow temperature range. The explanation usually given for this extreme slowing down of most of the molecular dynamics is that in such cases the thermal energy kT becomes similar to or smaller than the intermolecular potential energy required by the molecules to accommodate in the respective equilibrium configuration. Thus, the source of the high viscosity is the freezing of intramolecular configurations, while at sufficiently high temperatures the molecules are able to move past each other, allowing local stresses to be relieved at a much faster rate. In the case of nonrigid molecules such as most polymers this process is supported by the adaptation of the molecule to the constraints produced by the environment and external forces. In the high-viscosity state, on the other hand intramolecular barriers may prevent the molecule from undergoing configurational changes, this process leading to an increasing rigidity of the molecule itself. As the relaxation of the undercooled liquid to thermodynamic equilibrium becomes slower than the cooling rate, instead of crystal formation we observe the formation of a rigid amorphous glass. The temperature T_g at which this occurs is known as the glass point. Several authors define the glass point as the temperature at which

the rate of molecular motions pertinent to the relaxation to equilibrium becomes macroscopic (i.e., of the order of seconds to hours). Another definition stresses the aspect that glass formation can be viewed as a thermodynamic phase transition.

Macroscopically, glasses can be distinguished from ordinary liquids by the presence of elasticity. We express this by saying that glasses respond to external stress (mainly shear stress) predominantly by an elastic mechanism (full recovery after the stress has been relieved) while liquids respond by a viscous mechanism (no recovery after the stress has been relieved). Actually, both states of matter display viscoelasticity (i.e., viscous as well as elastic response); however, this is generally observed only in the intermediate cases where the response changes from predominantly viscous to predominantly elastic. The plot of the elasticity modulus versus temperature in Fig. 6 shows the transition from the glassy to the rubbery and from there to the liquid state.

The observation of viscoelastic behavior is, of course, a matter of the time scale of the experimental technique used to study the dynamics. Thus, the elastic response is apparent only when the time scale of the deformation is comparable to the time required for molecules to accommodate in the new equilibrium configuration. Viscoelastic properties of liquids and glasses are studied by measuring mechanical, ultrasonic, and rheological quantities such as various elastic moduli and viscosity coefficients. Furthermore, several spectroscopic techniques such as dielectric relaxation, time-dependent Kerr effect, and light-scattering spectroscopy have been applied successfully to the study of glass-forming systems. By these methods we obtain a characteristic relaxation time resulting from an exponential decay of some property such as dielectric polarization or from mechanical deformation. In many cases, however, the data indicate the presence of more than one relaxation process, which have been often described in terms of a slow α-process, a faster β-process, and other processes indicated by the Greek letters γ, δ, and so on.

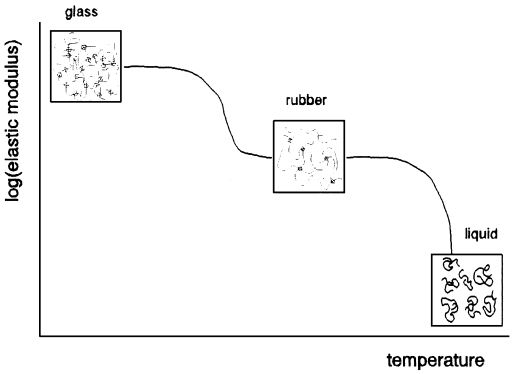


FIGURE 6 Plot of the elasticity modulus versus temperature, showing the transition from the glassy to the rubbery to the liquid state.

Especially the α-process is crucial in determining the mechanical properties of glass-forming systems, and attempts are made to synthesize molecules with a given temperature dependence of the α-relaxation process. This would allow us to obtain materials with definite useful mechanical properties in a given temperature range. The interpretation of these processes at a molecular level, however, is a still unsolved problem. Many attempts to rationalize the data in a semiphenomenological manner are based on the concept of the free volume available to the molecular motion and hence to the relaxation of nonequilibrium configurations. All these phenomena, which have been extensively studied in polymer melts, are also observed in several glass-forming low-molecular-weight liquids (e.g., o-terphenyl, decalin, salol, and polyalcohols).

An important observation in supercooled liquids is the dependence of the physical properties of these substances on the thermal history of the sample. The question whether glass formation is the effect of kinetic constraints only, or whether other factors play a role, is still open. The existence of metallic glasses and the observation of a glassy phase in computer simulations of molecules as simple as argon may be important clues to this question.

VII. GELS

When a low-molecular-weight liquid is dissolved in a high-molecular-weight system (the stationary component), which often is a cross-linked polymer, under certain conditions we observe the formation of a gel. This is a macroscopically homogeneous liquid with high internal mobility but no macroscopic steady-state flow. Gel formation requires the presence of more or less stable cross-links to prevent viscous flow as well as a fluid component that must be a good solvent for the stationary component.

Figure 7 displays schematically the structure of a gel.

VIII. DYNAMICS IN COMPLEX LIQUIDS

One of the dynamical problems studied theoretically and experimentally very extensively is the rotational motion of molecules in liquids. The molecular rotation in gases is solely determined by the kinetic energy and the moment(s) of inertia of the molecule. Under the action of frequent random collisions with other molecules, the rotation of a particular molecule in a liquid is continuously perturbed, and this is reflected in an exponential time dependence of the correlation function of the orientation of the molecular axes. Generally, at short times of the order of 0.1 ps and less, the motion is determined by the molecular

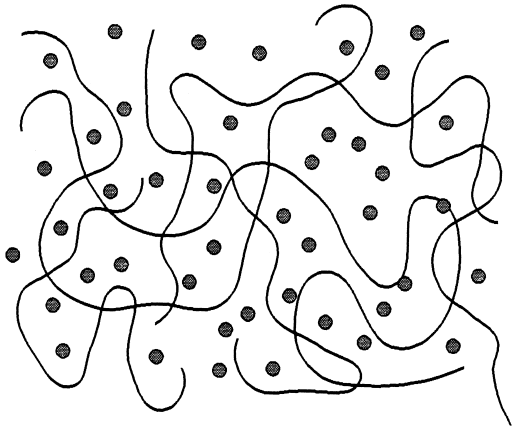


FIGURE 7 The structure of a gel. The points represent the molecules of the liquid. The cross-linked polymer is represented by the lines.

moment of inertia and the temperature, whereas at longer times, the motion is determined by angular momentum exchange due to the frequent collisions with other molecules. This can be described by a friction exerted on the rotating molecule by its neighbors. It was shown by Debye that under certain simplifying assumptions this so-called rotational diffusion can be described by a relaxation time τ_{OR} which is connected to the macroscopic viscosity η of the liquid:

$$\tau_{OR} = \frac{\eta V}{kT}.$$

In this equation V is a characteristic volume, the hydrodynamic volume of the molecule. If the shape and size of the rotating molecule are known, this relation can be used to probe the local viscosity in liquid systems (e.g., in micelles and membranes). Such a local viscosity can be different from the macroscopic viscosity and is accessible only through measurements done on label molecules.

Since molecular labels are convenient indicators of the local microdynamics of the liquid in their neighborhood, they can also be used to test theoretical models of liquid-state dynamics. The experimental methods currently used are NMR relaxation, Raman linewidth measurements, dynamic light-scattering spectroscopy, fluorescence anisotropy, and dielectric relaxation. The theory of rotating molecules in a liquid medium interacting in an uncorrelated random fashion with the surroundings has been described by models amenable to analytical calculations in the case of simple liquids. The quantities that enter the calculation are molecular moments of inertia, molecular masses, and intermolecular forces. In the case of more complex liquids, the assumption of a diffusive motion in a continuum is made, and the parameters of the model are hydrodynamic quantities that can be compared with the corresponding macroscopic data (e.g., the macroscopic shear viscosity).

The translational motion of very large molecules in liquids, such as diluted polymers or supramolecular aggregates like micelles and microemulsions, has been studied by light-scattering methods to obtain information about the molecular weight and size of the diffusing entity, its polydispersity, the interactions with other species (e.g., ions) or, at higher concentrations, interactions between the diffusing molecules themselves, and, finally, the internal flexibility and the rate of configurational changes.

IX. CONCLUSIONS

The liquid state includes a large number of phenomenologically very different systems such as simple liquids, micelles, microemulsions, polymer melts, liquid crystals, and gels. All these systems have in common (1) a rather high molecular mobility which may, however, be restricted in different ways depending on the system, and (2) molecular disorder which may also be restricted in different ways. The study of the liquid state involves most of the modern physical methods, and the theory of its molecular aspect requires elaborate statistical mechanical methods. The study of the liquid state is progressing at a rapid rate although several basic problems still remain unanswered.

SEE ALSO THE FOLLOWING ARTICLES

FLUID DYNAMICS • GLASS • HYDROGEN BOND • LIQUID CRYSTALS (PHYSICS) • MICELLES • MOLECULAR HYDRODYNAMICS • PERMITTIVITY OF LIQUIDS • POTENTIAL ENERGY SURFACES • RHEOLOGY OF POLYMERIC LIQUIDS • X-RAY SMALL-ANGLE SCATTERING

BIBLIOGRAPHY

Barnes, A. J., Orville-Thomas, W. J., and Yarwood, J., eds. (1983). "Molecular Dynamics and Interactions," D. Reidel, Dordrecht.

Berne, B. J., and Pecora, R. (1976). "Dynamic Light Scattering," Wiley, New York.

Birnbaum, G. (1985). "Phenomena Induced by Intermolecular Interactions," Plenum, New York.

Enderby, J. E., and Barnes, A. C. (1990). *Reports on Progress in Physics* **53**(1 & 2), 85–180.

Hansen, J. P., and McDonald, I. R. (1976). "Theory of Simple Liquids," Academic Press, New York.

Rothschild, W. G. (1984). "Dynamics of Molecular Liquids," Wiley, New York.

Rowlinson, J. S. (1982). "Liquids and Liquid Mixtures," Butterworth, London.

Wang, C. H. (1985). "Spectroscopy of Condensed Media," Academic Press, Orlando.



Mechanics, Classical

A. Douglas Davis
Eastern Illinois University

I. Kinematics
II. Newton's Laws of Motion
III. Applications
IV. Work and Energy
V. Momentum
VI. Rigid Body Motion
VII. Central Forces
VIII. Alternate Forms

GLOSSARY

Conservation Certain quantities—e.g., energy and momentum—remain the same for a system before, during, and after some interaction (often a collision). Such quantities are said to be conserved.
Dynamics Explanation of the cause of motion. This involves forces acting on massive bodies and the motion that ensues.
Energy Ability to do work; stored-up work.
Kinematics Description of motion.
Momentum Mass multiplied by the velocity. Momentum is a vector.
Statics Study of forces acting on bodies at rest.
Work Distance a body moves multiplied by the component of force in the direction of the motion. Work is a scalar.

CLASSICAL MECHANICS is the study of ordinary, massive objects, for example, the study of objects roughly the size of a bread box traveling at roughly sixty miles an hour. It is to be distinguished from quantum mechanics, which deals with particles or systems of particles that are extremely small, and it should also be distinguished from relativity, which deals with extremely high velocities.

Classical mechanics can be divided into statics, kinematics, and dynamics. Statics is the study of forces on a body at rest. Kinematics develops equations that merely describe the motion without question to its cause. Dynamics seeks to explain the cause of the motion.

I. KINEMATICS

Motion of an object must be described in terms of its *position* relative to some reference frame. If the motion occurs in one dimension (as along a straight highway or railroad track) the position will usually be written as x. If the motion occurs in two or three dimensions (as an airplane circling an airport or a spacecraft on its way to Jupiter) the position will be written as $\mathbf{r}$. The position is a

vector. For the one-dimensional case, the vector nature of the position shows up as the *sign* of x. For example, the position may be considered positive to the right; then it will be negative to the left. Position is commonly measured in meters. Of course, position may also be measured in kilometers, centimeters, feet, or miles as the need arises.

Velocity is the time rate of change of the position of an object. It can be written as

$$v = \Delta x/\Delta t$$

or

$$v = dx/dt$$

for the one-dimensional case or, for the three-dimensional case, as

$$\mathbf{v} = \Delta \mathbf{r}/\Delta t$$

or

$$\mathbf{v} = d\mathbf{r}/dt,$$

where x or $\mathbf{r}$ is the position of the object of interest. Velocity describes how fast the object is moving and in which direction. That means that velocity is a vector. Speed is the magnitude (just the "how fast" without the direction) of velocity. Speed is a scalar. Both are commonly measured in m/s.

Acceleration is the time rate of change of the velocity of an object. It can be written as

$$a = \Delta v/\Delta t$$

or

$$a = dv/dt$$

for the one-dimensional case or, for the three-dimensional case, as

$$\mathbf{a} = \Delta \mathbf{v}/\Delta t$$

or

$$\mathbf{a} = d\mathbf{v}/dt.$$

Acceleration is commonly measured in meters per second per second (m/s^2). An acceleration of 10 m/s^2 means that the velocity increases by 10 m/s every second. Other systems of units could be used. For example, automotive engineers may find it useful to express a car's acceleration in miles/h/s. An acceleration of 4.3 miles/h/s means that a car's velocity increases by 4.3 miles/h every second.

A. Constant Acceleration

If the position is known as a function of time, then the velocity and acceleration are quite easy to determine by applying their definitions. However, it is more usually the case that the acceleration is known and the velocity and position are wanted. For constant acceleration, a, in one dimension the velocity and position at some time t can be found from

$$v = v_0 + at$$

$$x = x_0 + v_0 t + \tfrac{1}{2}at^2,$$

where x_0 is the initial position at $t = 0$ and v_0 is the initial velocity at $t = 0$. Often it is useful to determine the velocity at some position, rather than at some time. These two equations can be solved to provide

$$v^2 = v_0^2 + 2a(x - x_0).$$

B. Nonconstant Acceleration

Acceleration is connected to position through a second-order differential equation

$$a = \frac{d^2x}{dt^2}$$

or

$$\mathbf{a} = \frac{d^2\mathbf{r}}{dt^2}$$

so the solution of $x(t)$ or $\mathbf{r}(t)$ from a or $\mathbf{a}$ may, indeed, be rather difficult. If the acceleration is known as a function of time, $a(t)$, then it may be integrated directly to yield

$$v(t) = v_0 + \int_0^t a(t)\,dt$$

and

$$x(t) = x_0 + \int_0^t v(t)\,dt.$$

A few other, special cases exist, which may be solved directly. In general, though, ideas from dynamics, such as energy conservation or momentum conservation, are usually necessary in solving for or understanding the motion of an object when its acceleration is not constant.

II. NEWTON'S LAWS OF MOTION

A. Inertia

Once a car is moving a braking system is necessary to bring it back to a stop. Or a book lying on a desk requires a push or a shove from the outside to start it moving. Both of these situations are examples of *inertia*. Because of inertia, an object tends to continue to do what it is presently doing. This seems to have been first understood by Galileo and was first clearly stated by Sir Isaac Newton in the first of his three laws of motion."In the absence of forces from the outside, a body at rest will remain at rest and a body in motion will continue in motion along the same straight line with the same velocity."

Friction is a force, which is nearly always present and sometimes masks this idea of inertia. If a book is given a shove across a table it may stop before reaching the edge. The law of inertia, Newton's first law of motion, is still valid. But there is a force from the outside, the force of friction. An ordinary car will not coast forever; it will eventually come to rest. But there are forces from the outside which cause this (namely friction due to the air and friction between tires and roadway.)

The idea of inertia is important because it asserts that motion continues because of the motion present. There need not be an active, continuing agent present at all times as the motion continues.

B. Force ($F = ma$)

While inertia is important, motion is far more interesting when there is a force present. If there are several forces acting on a body, it is the *net* force—the vector sum of all the forces—that is important. Newton's second law of motion states that "in the presence of a net or unbalanced force a body will experience an *acceleration*. That acceleration is inversely proportional to the *mass* of the body and is directly proportional to the *force* and in the direction of the force."

This can be written as $\mathbf{a} = \mathbf{F}/m$ although it is more commonly written as $\mathbf{F} = m\mathbf{a}$. There is little exaggeration to say that almost all of classical mechanics derives directly from Newton's second law.

Velocity and acceleration are easily and often confused. Most people are more familiar with velocity or speed. However, it is the acceleration that is of most use in determining and describing motion and its cause.

A force is a push or a pull. Force is anything that causes an acceleration. Newton's second law can be used as the definition of a force.

Newton's second law also provides an operational definition of the mass of an object. It is a measure of how much "stuff" there is in an object. It is a measure of how difficult it is to accelerate an object. By definition, a particular block of platinum–iridium alloy has been designated to have a mass of exactly 1 kg. If the same force is applied to this (or an identical) block and to another block and the other block is found to have an acceleration exactly one-half that of the standard block, then the other block's mass is 2 kg.

The mass of an object is always the same. It is independent of altitude or position. As we shall see, there is an important distinction between mass and weight.

In the metric system (or SI units), mass is measured in kilograms, force in newtons, and acceleration in m/s^2. A force of one newton could cause a 1 kg mass to accelerate at 1 m/s^2.

C. Action–Reaction

Newton's third law of motion states that "if object 1 exerts a force on object 2 then object 2 also exerts a force back on object 1. The two forces are identical in magnitude and opposite in direction."

An example of this is the force you exert down on a chair when you sit on it and the force the chair exerts up on you. When an airplane propeller pushes back on the air, the air pushes forward on the propeller. As the sun pulls on earth, earth also pulls back on the sun. It is impossible to exert a force on an object without an additional force being exerted by that object. Notice that the forces in question are always exerted on different objects.

III. APPLICATIONS

A. Straight-Line Motion

Any constant force produces a constant acceleration so the kinematic equations for constant acceleration are immediately useable.

1. Free Fall

Freely falling objects near the earth's surface are found to have a constant acceleration of 9.8 m/s^2 (or 32 ft/s^2) *downward* if air resistance can be neglected. This acceleration is usually labeled "g"; that is, $g = 9.8$ m/s^2 $= 32$ ft/s^2. To produce the same acceleration, the forces on two different bodies must be proportional to the masses. That means that the force of gravity must be proportional to the mass of a body. This force of gravity is called *weight* W and

$$W = mg.$$

The kinematics equations that describe an object in free fall, then, are simply

$$v = v_0 - gt$$
$$x = x_0 + v_0 t - \tfrac{1}{2} g t^2,$$

where the acceleration a has just been replaced with $-g$ (the minus sign merely indicates downward).

2. Simple Harmonic Motion

A spring exerts a linear restoring force. As a spring is stretched or compressed, the force it exerts is proportional to how far it has been stretched or compressed from its unstretched, uncompressed, equilibrium position. And the force is directed to move the stretched or compressed spring back to that equilibrium position. This force can be described by the equation

$$F = -kx,$$

where x is the displacement from equilibrium (positive for stretch and negative for compression), k is a spring constant that describes the strength of the spring, and F is the force.

If an object of mass m is attached to such a spring the motion that it undergoes is known as simple harmonic motion. That motion can be described by

$$x(t) = A\sin(\omega t + \varphi).$$

A is called the amplitude of the motion and is the maximum displacement from the equilibrium position. The motion is symmetric; the object will move as far on one side of the equilibrium position as on the other side. φ is a phase angle determined by the initial conditions (x_0, v_0). ω is the *angular frequency* in rad/s. It is related to the more usual frequency f of cycles per second by

$$f = 2\pi\omega.$$

For such a mass on a spring the angular frequency is equal to

$$\omega = \sqrt{k/m}.$$

The period T is the amount of time required for a single cycle; therefore, the period and frequency are related by

$$f = 1/T.$$

A small object of mass m suspended by a cord of length l and allowed to swing back and forth is called a simple pendulum. For small amplitudes the motion of such a simple pendulum is also simple harmonic motion. For the simple pendulum the angular frequency is given by

$$\omega = \sqrt{g/l}.$$

Note that for both examples of simple harmonic motion, the frequency is independent of the amplitude.

B. Three-Dimensional Motion

Newton's second law makes it easy to extend the ideas of straight-line motion to projectile motion, the motion followed by a body thrown and released near the earth's surface. Observe this motion from far, far away in the plane of the motion and it looks like the object has simply been thrown upward. The force of gravity acts to accelerate it downward. The vertical part of the motion appears to be simply free fall and that is just motion with constant acceleration. Observe this motion from far above and it looks like the object is moving at constant velocity. There is no horizontal force. The horizontal part of the motion appears to be simply constant velocity.

The path a projectile takes is a parabola. The range is the horizontal distance an object will go if it is thrown from and lands back on the same level surface. The range R is given by

$$R = v_0^2 \sin 2\theta / g,$$

where v_0 is the initial speed and θ is the angle above the horizontal at which the projectile is thrown. Note that the range is the same for complimentary angles; that is, θ and $90^\circ - \theta$ give the same range. Maximum range is found for $\theta = 45^\circ$.

IV. WORK AND ENERGY

A. Work

Work done by a constant force F is defined as the distance D an object moves multiplied by the component of force in that direction. Pushing on a wall may tire your body but no work has been done according to this definition. If a yo-yo swings in a circle, the string continually exerts a force perpendicular to the direction of motion and no work is done. Work is a scalar quantity. The units of work are newton-meters or joules.

B. Kinetic Energy

The amount of work done on a body is equal to an increase in the quantity $\frac{1}{2}mv^2$. That is,

$$W = \tfrac{1}{2}mv_f^2 - \tfrac{1}{2}mv_0^2,$$

where v_f is the final speed after the work has been done and v_0 was the original speed before. Because the object is moving, it has the ability to do work on something else—it could exert a force on another object over some distance. This ability to do work is called energy. Energy is a scalar. The quantity $\frac{1}{2}mv^2$ is called the kinetic energy; it is energy due to motion.

The kinetic energy associated with the random motion of molecules due to heat is called thermal energy.

C. Potential Energy

Doing work on an object may change its position or condition. Lifting an object requires doing work against gravity. Because of its higher position, the object can then do work on something else as it falls; thus, it has gravitational potential energy. If an object of mass m is lifted from an initial height y_0 to a final height y its potential energy is changed by an amount

$$\Delta PE = mg(y - y_0).$$

Stretching or compressing a spring requires work to be done. That work done is stored up in the spring; the spring can be released and can do work on something

else. This elastic potential energy of a spring stretched or compressed a distance x from its equilibrium position is

$$PE = \tfrac{1}{2}kx^2.$$

D. Conservation of Energy

Work and energy are useful because the work done on a system by forces from outside the system is equal to the change in the total energy of the system. The total energy of a system is the sum of the potential, kinetic, and thermal energies. If the work from external forces is zero then the total energy of the system remains constant—the total energy is conserved.

A simple pendulum is an example of a system for which the external forces do no work. The force exerted by the supporting string on the mass of a pendulum is always perpendicular to the direction of motion so no work is done. Therefore, the energy must be conserved. If a pendulum is lifted some distance and released, it begins with some amount of gravitational potential energy. As it swings that potential energy decreases but its speed increases, which means the kinetic energy increases. The sum of the kinetic and potential energies remains constant.

A roller coaster offers another example of a system that alternately changes potential energy (height) into kinetic energy (speed) and vice versa. For both a roller coaster and a pendulum, friction will eventually cause the system to stop. Friction can be considered an external force or we can look at the thermal energy associated with the slight increase in temperature of the wheels and rails as a roller coaster runs.

V. MOMENTUM

Momentum, usually designated by $\mathbf{p}$, is defined by multiplying the mass m of an object by its velocity $\mathbf{v}$,

$$\mathbf{p} = m\mathbf{v}.$$

It is similar to kinetic energy in that momentum increases with increasing speed. But it is different in that momentum is a vector quantity.

Like energy, momentum is useful because it is conserved. In the absence of external forces, the total momentum of a system of particles remains constant. Even though the internal forces between the particles may be very complicated, the vector sum of all the momenta of all the particles remains constant. Conservation of momentum is related to Newton's third law of motion (action and reaction).

A. Collisions

When two objects collide—as two billiard balls hitting or two cars crashing into each other—the forces are very difficult to measure or predict. But conservation of momentum means that the vector sum of the momenta of the two objects before the collision will be the same as the vector sum of the momenta of the two objects after the collision. By itself, this is not sufficient to completely solve for the velocities of the two objects after the collision (assuming the conditions before the collision are given). But the final velocities can be found in two very useful extremes. If the kinetic energy is also conserved, that is, no energy is lost to heat or deforming the objects, the collision is termed *elastic*. The additional information provided is enough to solve for the final motion. If the two objects stick together the collision is termed *inelastic* and the maximum amount of kinetic energy is lost. Note that momentum is always conserved whether the collision is totally elastic, totally inelastic, or anywhere in between.

B. Rocket Propulsion

A car's motion can be understood by looking at the wheels as they push on the pavement and understanding that the pavement pushes back on the wheels. But how, then, does a rocket move and accelerate in space? There is nothing else around for it to push on that can push back on it. A rocket burns fuel that is exhausted from the rocket's engine at high velocity. As momentum is carried in one direction by the fuel, an equal amount of momentum is carried in the opposite direction by the rocket. If you stand in a child's wagon and throw bricks in one direction you will be moved in the other direction. As momentum is carried in one direction by the bricks, an equal amount of momentum is carried in the opposite direction by you and the wagon. The idea is the same as that used in explaining rocket propulsion. If gravity can be neglected, a rocket's final velocity is given by

$$v = v_0 + u \ln (m_0/m),$$

where v_0 is its initial velocity, u is the exhaust velocity of the burned gases, m_0 is the initial mass of the rocket, and m is the final mass of the rocket.

VI. RIGID BODY MOTION

A. Center of Mass

For a system of particles of mass m_i each located at position $\mathbf{r}_i$, the mass-weighted average position of the particles is called the center of mass and is defined by

$$\mathbf{R} = \left(\sum_i m_i \mathbf{r}_i\right) \Big/ \left(\sum_i m_i\right),$$

where $\sum_i$ means to sum over all values of i. The total mass of the system of particles is $M = \sum_i m_i$.

For a rigid body the summation over individual masses is replaced by an integral over the volume of the body. The center of mass is then defined by

$$\mathbf{R} = \frac{1}{M}\iiint_V \rho \mathbf{r}\, dV,$$

where M is the total mass of the body, given by

$$M = \iiint_V \rho\, dV.$$

ρ is the mass density (mass per unit volume), $\mathbf{r}$ is just the location vector, and V is the volume of the body. As with all vector equations, this may be easier to understand in component form. The three coordinates (X, Y, Z) of the center of mass are

$$X = \frac{1}{M}\iiint_V \rho x\, dV,$$

$$Y = \frac{1}{M}\iiint_V \rho y\, dV,$$

$$Z = \frac{1}{M}\iiint_V \rho z\, dV.$$

The center of mass is a "uniquely interesting point" for even though the motion of individual particles or rotations of the body may be frustratingly complicated, the motion of the center of mass will be that of a single point particle with mass M.

B. Angular Momentum

Just as linear momentum was useful in understanding and predicting translational motion because of its conservation, so another conserved quantity (called the angular momentum) will be useful in discussing rotational motions. The angular momentum $\mathbf{L}$ of a small particle relative to some origin is given by

$$\mathbf{L} = \mathbf{r} \times \mathbf{p},$$

where $\mathbf{r}$ is the location of the particle from the origin, $\mathbf{p}$ is its momentum, and $\times$ indicates the vector cross product. For a system of particles, the total angular momentum is the vector sum of the individual angular momenta. For an extended body, the total angular momentum requires evaluating an integral over the volume of the body.

1. Rotation about a Fixed Axis

For rotation about a fixed axis, there is a strong correlation with straight-line motion. The mass is replaced by a "rotational mass" that depends upon the geometry of the mass (how far it is located from the axis of rotation.) This "rotational mass" is called the moment of inertia I. For a hollow cylinder of mass M and radius R, the moment of inertia is $I = MR^2$. For a solid cylinder, $I = \frac{1}{2}MR^2$. Force is replaced by a "rotational force" that depends upon the force and its placement from the axis of rotation; this is called a torque T. While a small force applied at the doorknob side opens a door easily, a large force will be required if it is applied back near the hinge; the rotational effect in the two cases is the same. Torque is given by

$$T = rF\sin\theta,$$

where r is the distance from the axis of rotation, F is the force, and θ is the angle between the two.

Just as a distance x labels the position of a mass on a straight track, an angle θ (measured in radians) labels the angular position of a rotating object. Angular velocity ω describes its speed of rotation in rad/s and angular acceleration α describes the rate of change of angular velocity in rad/s^2.

The rotational equivalent of $F = ma$ is $T = I\alpha$. The angular momentum for rotation about a fixed axis is $L = I\omega$, which closely parallels $P = Mv$ for the linear case.

2. Rotation in General

In general, however, rotation can be more complicated than straight-line motion. Angular momentum remains a conserved quantity. But in general angular momentum is given by

$$\mathbf{L} = \{\mathbf{I}\}\boldsymbol{\omega},$$

where $\{\mathbf{I}\}$ is now a tensor. This brings about the interesting case in which the angular momentum $\mathbf{L}$ and the angular velocity $\boldsymbol{\omega}$ may not necessarily be parallel to each other. This can be seen by tossing a book or tennis racket in the air spinning about each of three mutually perpendicular axes. For the longest and shortest axes, $\mathbf{L}$ and $\boldsymbol{\omega}$ will be in the same direction; for the medium length axis they will not be in the same direction.

VII. CENTRAL FORCES

A. Definitions

A central force is one whose direction is always along a radius; that is, either toward or away from a point that can be used as an origin (or force center), and whose magnitude depends solely upon the distance from that origin, r. A central force can always be written as

$$\mathbf{F} = F(r)\hat{\mathbf{r}},$$

where $\hat{\mathbf{r}}$ is a unit vector in the radial direction. Central forces are important because many real situations involve central forces. The gravitational force between two masses

and the electrostatic force between two charges are both central forces. Motion due to a central force will always be confined to a plane.

B. Gravity

Gravity is the force of attraction between two massive bodies. First described by Sir Isaac Newton, the force of gravity between two bodies with masses m_1 and m_2 separated by a distance r is given by

$$F_G = (Gm_1m_2)/r^2,$$

where G is a universal constant ($G = 6.672 \times 10^{-11}$ N m^2/kg^2). This expression is valid for calculating the force earth exerts on an apple near its surface or the force earth exerts on our moon or the force our sun exerts on Jupiter.

1. Kepler's Laws of Planetary Motion

Before Newton discovered this law of universal gravitation, Johannes Kepler found, based upon careful observational data, that the motion of the planets in our solar system could be explained by three laws:

1. Planets move in orbits that are ellipses with the sun at one focus (elliptical orbits).
2. Areas swept out by the radius vector from the sun to a planet in equal times are equal (equal areas in equal times).
3. The *square* of a planet's period is proportional to the *cube* of the semimajor axis of its orbit ($T^2 \propto r^3$).

It was a great triumph of Newton's law of universal gravitation that it could explain and predict Kepler's laws of planetary motion. Kepler's second law is true for any central force; it is the result of conservation of angular momentum. The other two laws depend upon gravity being an inverse square force.

2. Orbits

Planets travel in elliptical orbits about the sun. Satellites travel in elliptical orbits about their planet. If the speed of a satellite is suddenly increased the shape of the elliptical orbit elongates. If a satellite has enough velocity to escape and never return to the planet the path it travels is a parabola or a hyperbola. Escape velocity is the minimum velocity that will allow a satellite to travel away from its planet and never return. If a satellite leaves earth's surface with a velocity of about 40,000 km/h (25,000 miles/h) it will escape from earth and never return.

C. Harmonic Oscillator

A mass suspended between three sets of identical, mutually perpendicular springs forms an isotropic, three-dimensional simple harmonic oscillator. The springs provide a restoring force of the form

$$\mathbf{F} = -k\mathbf{r}$$

so this three-dimensional harmonic oscillator experiences a central force. Examples of such a system are atoms in certain crystals, where the interatomic bonds act as the springs in this simple case. If the springs (or interatomic bonds for a crystal) are not all identical, then the force due to a displacement in one direction will be different than that for another direction. The harmonic oscillator is anisotropic and can no longer be described as a central force.

VIII. ALTERNATE FORMS

Newton's second law of motion, $F = ma$, can be used to solve for the motion in many situations. But the same information can be written in different forms and used in situations where direct solution of $F = ma$ is very difficult or perhaps impossible.

A. Lagrange's Equations

Lagrange's equations of motion can be written as

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} = \frac{\partial L}{\partial q_k},$$

where q_k is a "generalized coordinate" and L is called the Lagrangian function. The Lagrangian function is the difference between the kinetic energy and the potential energy; $L = KE - PE$. The dot means a time derivative; $\dot{q}_k = dq_k/dt$.

B. Hamilton's Equation

Hamilton's equations of motion can be written as

$$\dot{q}_k = \partial H/\partial p_k$$

and

$$\dot{p}_k = -\partial H/\partial q_k,$$

where, again, q_k is a "generalized coordinate," p_k is a "generalized momentum," and H is called the Hamiltonian function. For many situations, the Hamiltonian H is the total energy of the system.

C. Poisson Brackets and Quantum Mechanics

Hamilton's equations can be rewritten in terms of Poisson brackets as

$$\dot{q}_k = [q_k, H]$$

and

$$\dot{p}k = [p_k, H],$$

where the Poisson brackets are defined by

$$[A, B] = \sum_k \left[\frac{\partial A}{\partial q_k} \frac{\partial B}{\partial p_k} - \frac{\partial B}{\partial q_k} \frac{\partial A}{\partial p_k} \right].$$

This formulation is especially interesting because it allows for an easy and direct transfer of ideas from classical mechanics to quantum mechanics.

SEE ALSO THE FOLLOWING ARTICLES

CELESTIAL MECHANICS • CRITICAL DATA IN PHYSICS AND CHEMISTRY • ELECTROMAGNETICS • MECHANICS OF STRUCTURES • NONLINEAR DYNAMICS • QUANTUM MECHANICS • RELATIVITY, GENERAL • STATISTICAL MECHANICS • VIBRATION, MECHANICAL

BIBLIOGRAPHY

Arya, A. P. (1997). "Introduction to Classical Mechanics," Prentice Hall, New York.

Kwatny, H. G., and Blankenship, G. L. (2000). "Nonlinear Control and Analytical Mechanics: Computational Approach," Birkhauser, Boston.

Brumberg, V. A. (1995). "Analytical Techniques of Celestial Mechanics," Springer-Verlag, Berlin.

Chow, T. L. (1995). "Classical Mechanics," Wiley, New York.

Doghri, I. (2000). "Mechanics of Deformable Solids: Linear and Nonlinear, Analytical and Computational Aspects," Springer-Verlag, Berlin.

Hand, L. N., and Finch, J. D. (1998). "Analytical Mechanics," Cambridge University Press, Cambridge, UK.

José, J. V., and Saletan, E. J. (1998). "Classical Dynamics: A Contemporary Approach," Cambridge University Press, Cambridge, UK.

Torok, J. S. (1999). "Analytical Mechanics: With an Introduction to Dynamical Systems," Wiley, New York.



Molecular Hydrodynamics

Sidney Yip
Massachusetts Institute of Technology

Jean Pierre Boon
Université Libre de Bwxelles

I. Motivation
II. Density Correlation Function
III. Linearized Hydrodynamics
IV. Generalized Hydrodynamics
V. Kinetic Theory
VI. Mode Coupling Theory
VII. Lattice Gas Hydrodynamics

GLOSSARY

Diffusion Dissipation of thermal fluctuations by essentially random (or stochastic) motions of atoms (as in thermal or concentration diffusion).

Generalized hydrodynamics Theoretical description of fluctuations in fluids based on the extension of the equations of linearized hydrodynamics to finite frequencies and wavelengths.

(k, ω) space Region of wavenumber and frequency where thermal fluctuations are being studied.

Lattice gas automata Class of cellular automata designed to model fluid systems using discrete space and time implementation.

Memory function Space–time dependent kernel appearing in the equation of motion for time correlation function, which contains the effects of static and dynamical interactions.

Mode coupling A theory in which the interatomic interactions are expressed in terms of products of two or more modes of thermal fluctuations, such as the densities of particle number, current, and energy.

Propagation Cooperative motion of atoms characterized by a peak at finite frequency in the frequency spectrum of density fluctuations (as in pressure wave propagation).

Thermal fluctuations Spontaneous localized fluctuations in the particle number, momentum, and energy densities of atoms in a fluid at thermal equilibrium.

Time correlation function Function that expresses the correlation of dynamical variables evaluated at two different time (and space) points.

Uncorrelated binary collisions Sequence of two-body collisions in which the collisions are taken to be independent even though pairs of atoms can recollide one or more times.

MOLECULAR HYDRODYNAMICS is the theoretical description of spontaneous localized fluctuations in space and time of the particle number density, the current density, and the energy density in a fluid at thermal equilibrium. Its domain of applicability ranges from low frequencies and long wavelengths, where the linearized equations of

hydrodynamics are applicable, to frequencies and wavelengths comparable to interatomic collision frequencies and mean free paths.

I. MOTIVATION

When a fluid is disturbed locally from equilibrium, it will relax by allowing the perturbation to dissipate throughout the system. At the macroscopic level this response involves the processes of mass diffusion, viscous flow, and thermal conduction, which are the mechanisms by which the transport of mass, momentum, and energy can take place. In the absence of an external disturbance, we can still speak of the dynamical behavior of a fluid in these terms. The reason is that the same processes also govern the dissipation of spontaneous fluctuations that are always present on the microscopic level in a fluid at finite temperature. So the fluid can be considered as a "reservoir of thermal excitations" extending over a broad range of wavelengths and frequencies from the hydrodynamic scale down to the range of the intermolecular potential. Thus, the study of thermal fluctuations is fundamental to the understanding of the molecular basis of fluid dynamics.

The conventional theory of fluid dynamics invariably begins with the equations of hydrodynamics. The basic assumption of *hydrodynamics* is that changes in the fluid take place sufficiently slowly in space and time that the system can be considered to be in a state of local thermodynamic equilibrium. Under this condition we have a closed set of equations describing the space–time variations of the conserved variables, namely, the mass, momentum, and energy densities. These equations become explicit, when the thermodynamic derivatives and the transport coefficients occurring in them are known; however, such constants are not determined within the hydrodynamic theory, and therefore must be provided by either measurement or more fundamental calculations.

The equations of hydrodynamics have an extremely wide range of scientific and technological applications. They are valid for disturbances of arbitrary magnitude provided the space and time variations are slow on the molecular scales with lengths measured in collision mean free path l and times in inverse collision frequency ω_c^{-1}. In terms of the wavelength, $2\pi/k$, and frequency ω, of the fluctuations, the hydrodynamic description is valid only in the region of low (k, ω), where $kl \ll 1$ and $\omega \ll \omega_c$.

When the condition of slow variations is not fully satisfied, we expect the fluid behavior to show molecular or nonhydrodynamic effects. Unless the fluctuations are far removed from the hydrodynamic region of (k, ω), the discrepancies often appear only in a subtle and gradual manner. This suggests that extensions or generalizations of the hydrodynamic description may be useful and may be accomplished by retaining the basic structure of the equations, while replacing the thermodynamic derivatives and transport coefficients by functions that directly reflect the molecular structure of the fluid and the effects of individual intermolecular collisions. The result is then a theory that is valid even on the scales of collision mean free path and mean free time, a theory that may be called *molecular hydrodynamics.* In essence, molecular hydrodynamics is a description that considers both the macroscopic behavior of mass, momentum, and energy transport, and the microscopic properties of local structure and intermolecular collisions.

There are several reasons why a study of the extension of hydrodynamics is important. First, we obtain a better understanding of the validity of hydrodynamics. Second, an appreciation of how the details of molecular structure and collisional dynamics can affect the behavior of the conserved variables is essential to the study of transport phenomena on the molecular level. Finally, it is one of the basic aims of nonequilibrium statistical mechanics to develop a unified theory of liquids that treats not only the processes in the hydrodynamic region of (k, ω), but also the molecular behaviors that manifest at higher values of wavenumber and frequency.

II. DENSITY CORRELATION FUNCTION

The fundamental quantities in the study of thermal fluctuations in fluids are space and time-dependent correlation functions. These functions are the natural quantities for theoretical analyses as well as laboratory measurements. They are well defined for a wide variety of physical systems, and they possess both macroscopic properties and interpretations at the microscopic level.

For the fluid system of interest we imagine an assembly of N identical particles (molecules), each of mass m, contained in a volume Ω. The molecules have no internal degrees of freedom, and they are assumed to interact through a two-body, additive, central potential $u(r)$. The fluid is in thermal equilibrium, at a state far from any phase transition. Also, there are no external fields imposed, so the system is invariant to spatial translation, rotation, and inversion.

A time correlation function is the thermodynamic average of a product of two dynamical variables, each expressing the instantaneous deviation of a fluid property from its equilibrium value. The dynamical variables that we wish to consider are the number density,

$$n(\mathbf{r}, t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{R}_i(t)) \tag{1}$$

where $\mathbf{R}_i(t)$ denotes the position of particle i at time t, and the current density,

$$\mathbf{j}(\mathbf{r}, t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{v}_i(t)\delta(\mathbf{r} - \mathbf{R}_i(t)) \tag{2}$$

where $\mathbf{v}_i(t)$ is the velocity of particle i at time t.

The thermodynamic average of a dynamical variable $A(\mathbf{r}, t)$ is defined as

$$\langle A(\mathbf{r}, t)\rangle = \int d^3R_1 \ldots d^3R_N\, d^3P_1 \ldots \times\, d^3P_N\, f_{\text{eq}}(\mathbf{R}^N, \mathbf{P}^N)A(\mathbf{r}, t) \tag{3}$$

where f_{eq} is an equilibrium distribution of particle positions $\mathbf{R}^N = (\mathbf{R}_1, \ldots, \mathbf{R}_N)$, and momenta $\mathbf{P}^N = (\mathbf{P}_1 \ldots \mathbf{P}_N)$. Typically we adopt the canonical ensemble in evaluating Eq. (3),

$$f_{\text{eq}}(\mathbf{R}^N, \mathbf{P}^N) = Q_N^{-1} \exp(-\beta U) \prod_{i=1}^{N} f_0(\mathbf{P}_i) \tag{4}$$

with $\beta = (k_B T)^{-1}$, T being the fluid temperature and k_B the Boltzmann's constant. $U(\mathbf{R}^N)$ is the potential energy, $f_0(P)$ is the normalized Maxwell–Boltzmann distribution

$$f_0(P) = (\beta/2\pi m)^{3/2} \exp(-\beta P^2/2m) \tag{5}$$

and Q_N is the configurational integral

$$Q_N = \int d^3R_1 \ldots d^3R_N \exp(-\beta U) \tag{6}$$

where U is the potential energy of the system. Applying Eq. (3) to Eqs. (1) and (2) gives the average values $\langle n(\mathbf{r}, t)\rangle = \sqrt{N}/V$, and $\langle \mathbf{j}(\mathbf{r}, t)\rangle = 0$. Notice that in general a dynamical variable depends on the particle positions $\mathbf{R}^N$ and momenta $\mathbf{P}^N$, and also on the position $\mathbf{r}$ and time t, where the property is being considered. On the other hand, the average values are independent of $\mathbf{r}$ and t because the system is uniform and in equilibrium.

Given the dynamical variable $n(\mathbf{r}, t)$ we define the time-dependent density correlation function as

$$G(|\mathbf{r} - \mathbf{r}'|, t - t') = V\langle \delta n(\mathbf{r}', t')\delta n(\mathbf{r}, t)\rangle = \frac{1}{n}\left\langle \sum_{i,j}^{N} \delta(\mathbf{r}' - \mathbf{R}_i(t')) \times \delta(\mathbf{r} - \mathbf{R}_j(t))\right\rangle - n \tag{7}$$

where $\delta n(\mathbf{r}, t) = n(\mathbf{r}, t) - \langle n(\mathbf{r}, t)\rangle$, and $n = N/V$ is the average number density of the fluid at equilibrium. Despite its rather simple appearance this function contains all the structural and dynamical information concerning density fluctuations. Note that G depends on the separation $|\mathbf{r} - \mathbf{r}'|$ because of rotational invariance and it is a function of $t - t'$ because of time translational invariance. Without loss of generality we can take $\mathbf{r}' = 0$ and $t = 0$.

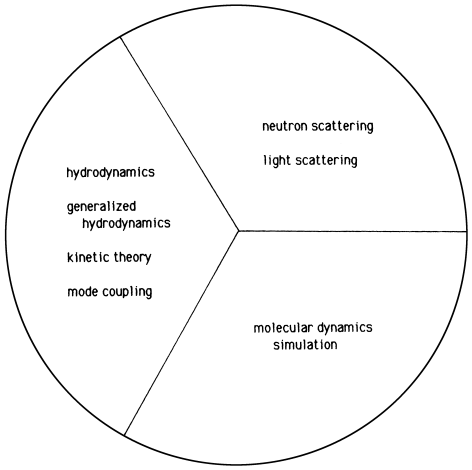


FIGURE 1 The time correlation function circle with its interconnected segments of theory, experiment, and atomistic simulation.

The density correlation function is the leading member of a group of time correlation functions that have received attention in recent studies of nonequilibrium statistical mechanics. These functions have become the standard language for experimentalists and theorists alike, because they can be measured directly and they are well-defined quantities for which microscopic calculations can be formulated. Moreover, time correlation functions are accessible by atomistic simulations. Figure 1 shows the complementary nature of theoretical, experimental, and simulation studies of time correlation functions. In this article we are primarily concerned with the theoretical developments, which, however, rely on simulation data and scattering measurements for guidance and validation.

It is instructive to note the simple physical interpretation of $G(r, t)$, which we can deduce from its definition. Consider a laboratory coordinate system placed in the fluid such that at time $t = 0$ a particle is at the origin. At a later time t, place an element of volume d^3r at the position $\mathbf{r}$. Then $G(\mathbf{r}, t)d^3r$ is the average (or expected) number of particles in the element of volume at $\mathbf{r}$ at time t, given that a particle was located at the origin initially. The initial value of G is

$$G(r, 0) = \delta(\mathbf{r}) + ng(r) \tag{8}$$

where $g(r)$ is the equilibrium pair distribution function

$$n^2 g(r) = \sum_{\substack{i,j \\ i\neq j}} \langle \delta(\mathbf{r} - \mathbf{R}_i)\delta(\mathbf{R}_j)\rangle \tag{9}$$

III. LINEARIZED HYDRODYNAMICS

The linearized hydrodynamic equations for a fluid with no internal degrees of freedom consist of the continuity equation, which expresses mass or particle number conservation,

$$\frac{\partial \rho_1(\mathbf{r}, t)}{\partial t} + \rho_0 \nabla \cdot \mathbf{v}(\mathbf{r}, t) = 0 \tag{10}$$

the Navier–Stokes equation, which expresses momentum or current conservation,

$$\rho_0 \frac{\partial \mathbf{v}(\mathbf{r}, t)}{\partial t} + \frac{c_0^2}{\gamma} \nabla \rho_1(\mathbf{r}, t) + \frac{c_0^2 \alpha \rho_0}{\gamma} \nabla T_1(\mathbf{r}, t) - \eta \nabla \nabla \cdot \mathbf{v}(\mathbf{r}, t) = 0 \tag{11}$$

and the energy transport equation, which expresses kinetic energy conservation,

$$\rho_0 C_v \frac{\partial T_1(\mathbf{r}, t)}{\partial t} - \frac{C_v(\gamma - 1)}{\alpha} \frac{\partial \rho_1(\mathbf{r}, t)}{\partial t} - \lambda \nabla^2 T_1(\mathbf{r}, t) = 0 \tag{12}$$

In these equations, $\rho = \rho_0 + \rho_1$ is the local number density, $T = T_0 + T_1$ is the local temperature, and $\mathbf{v}$ is the velocity, with subscripts 0 and 1 denoting the equilibrium value and instantaneous deviation, respectively. The ratio of specific heats at constant pressure and constant volume C_p/C_v is γ. The combination of shear and bulk viscosities $\frac{4}{3}\eta_s + \eta_B$ is denoted by η, and c_0, α, λ are, respectively, the adiabatic sound speed, the thermal expansion coefficient, and the thermal conductivity. Equations (10)–(12) are linearized in the sense that ρ_1, T_1, and $\mathbf{v}$ are assumed to be small, and therefore, only terms to first order in these quantities need be kept. This assumption makes the description valid only for small-amplitude disturbances such as thermal fluctuations.

The parameters of the hydrodynamic description are thermodynamic coefficients α, γ, and c_0, the thermal expansion coefficient, the ratio of specific heats at constant pressure and volume, and the adiabatic sound speed, and transport coefficients, η_s, η_B, and λ, the shear and bulk viscosities with $\eta = \frac{4}{3}\eta_s + \eta_B$, and the thermal conductivity. Once these are specified, the equations can be used to calculate explicitly the spatial and temporal distributions of the particle number, current, and energy densities for a given set of boundary and initial conditions.

We will be interested in the decay of a density pulse created by thermal fluctuations in a uniform, infinite fluid medium. For this problem it will be most convenient to discuss the solutions in wavevector space by taking the Fourier transform in configuration space and solving the resulting equations as an initial value problem with initial values,

$$\begin{aligned} \rho_1(\mathbf{r}, t = 0) &= \delta(\mathbf{r}) + n[g(r) - 1] \\ \mathbf{v}(\mathbf{r}, t = 0) &= 0 \\ T_1(\mathbf{r}, t = 0) &= 0 \end{aligned} \tag{13}$$

Equation (13) states that at time $t = 0$ a density pulse occurs in the fluid in the form of a particle localized at the origin of the coordinate system plus a distribution of particles according to $n[g(r) - 1]$; also, there are no current or temperature perturbations. The meaning of $\rho_1(\mathbf{r}, t)$ as the density response to this initial condition is the spatial distribution of this density pulse as time evolves. Notice that any particle in the fluid can contribute to $\rho_1(\mathbf{r}, t)$ for $t > 0$, not just the particle originally located at the origin. After taking the Fourier transform of Eqs. (10)–(12), we can solve for

$$n(\mathbf{k}, t) = \int d^3 r \, e^{i\mathbf{k}\cdot\mathbf{r}} \rho_1(\mathbf{r}, t) \tag{14}$$

The calculation is best carried out by taking the Laplace transform in time, for example,

$$n(k, s) = \int_0^\infty dt \, e^{-st} n(k, t) \tag{15}$$

thus obtaining a system of coupled algebraic equations for the Laplace–Fourier transformed densities $n(k, s)$, $\mathbf{v}(k, s)$, and $T_1(k, s)$. The system of equations is homogeneous, and for nontrivial solutions the transform variable s has to satisfy a cubic equation.

Since the hydrodynamics description is applicable only when spatial and temporal variations of the densities occur smoothly, it is appropriate to look for roots of the cubic equation to lowest orders in the wavenumbers. To order two,

$$\begin{aligned} s_\pm &= \pm i c_0 k - \Gamma k^2 \\ s_3 &= -\lambda k^2 / \rho_0 c_p \end{aligned} \tag{16}$$

where $\Gamma = [\eta + \lambda(C_v^{-1} - C_p^{-1})]/2\rho_0$. As a result of the density pulse, both pressure and temperature fluctuations are induced. The pair of complex roots $s_\pm$ describes the propagation of pressure fluctuations as damped sound waves, with speed c_0 and attenuation Γ. The root s_3 describes the diffusion of temperature fluctuations with attenuation $\lambda/\rho_0 C_p$.

Using Eq. (16) we can invert the Laplace transformed solution for $n(k, s)$ and compute the correlation function. The result is

$$\begin{aligned} F(k, t) &= \langle n(\mathbf{k}, t) n(-\mathbf{k}) \rangle \\ &= \int d^3 r e^{i\mathbf{k}\cdot\mathbf{r}} G(\mathbf{r}, t) \end{aligned} \tag{17}$$

$$F(k,t) = S(k)\left[\frac{C_p - C_v}{C_p}\exp\left(-\frac{\lambda k^2}{\rho_0 C_p}t\right) + \frac{C_v}{C_p}\exp(-\Gamma k^2 t)\cos c_0 k t\right] \quad (18)$$

where

$$S(k) = \langle n(\mathbf{k})n(-\mathbf{k})\rangle = F(k, t=0) \quad (19)$$

is known as the static structure factor of the fluid. In the long wavelength limit, it is a thermodynamic quantity: $S(k\to 0)\to nk_B T x_T$, where x_T is the isothermal compressibility. Equation (18) shows that there are two components in the time decay of density fluctuations, an exponential decay associated with heat diffusion, and a damped oscillatory decay associated with pressure (sound) propagation.

The dynamics of density fluctuations can be studied directly by scattering beams of thermal neutrons or laser light from the fluid and measuring the frequency spectrum of the scattered radiation. In such experiments, the frequency spectrum of the density fluctuation is measured,

$$S(k,\omega) = \int_{-\infty}^{\infty} dt\ e^{-i\omega t}F(k,t) \quad (20)$$

In contrast to $S(k)$, which is what we obtain from a neutron or X-ray diffraction experiment, $S(k,\omega)$ is called the dynamic structure factor because it gives information about both structure and dynamics of the fluid.

Since we can probe the fluid structure at different wavenumbers, the frequency behavior of $S(k,\omega)$ can vary considerably from the hydrodynamic regime of long wavelengths ($kl \ll 1$), to the regime of free particle flow ($kl \gg 1$). The frequency spectrum of density fluctuations in the hydrodynamic regime is characterized by three well-defined spectral lines, corresponding to the three modes in $F(k,t)$ or the three roots to the dispersion equation as given in Eq. (16). From (18) and (20), one obtains

$$S(k,\omega) = S(k)\left\{\frac{C_p - C_v}{C_p}\frac{\lambda k^2/\rho_0 C_p}{\omega^2 + \left(\lambda k^2/\rho_0 C_p\right)^2} + \frac{C_v}{C_p}\left[\frac{\Gamma k^2}{(\omega + c_0 k)^2 + (\Gamma k^2)^2} + \frac{\Gamma k^2}{(\omega - c_0 k)^2 + (\Gamma k^2)^2}\right]\right\} \quad (21)$$

The spectrum is composed of a central peak with maximum at $\omega = 0$ and whose full width at half maximum is $2\lambda k^2/\rho_0 C_p$. This peak is called the Rayleigh line; its intensity is given by $S(k)[1 - 1/\gamma]$. There are also two equally displaced side peaks with maxima at $\omega_\pm = \pm c_0 k$ and whose full width at half maximum is $2\Gamma k^2$; these are called the Brillouin doublet and their integrated intensity is given by $S(k)/\gamma$. The intensity ratio of the Rayleigh component to the Brillouin components is $\gamma - 1$, a quantity known as the Landau–Placzek ratio. Note that a more accurate solution contains cross terms involving heat diffusion and pressure propagation and gives rise to an asymmetry in the Brillouin components.

There are other time correlation functions of interest, such as the transverse current correlation,

$$J_t(k,t) = \frac{1}{N}\sum_{j,k}\left\langle v_j^T(t)v_k^T(0)e^{ik\cdot[\mathbf{R}_j(t)-\mathbf{R}_k(0)]}\right\rangle \quad (22)$$

where $v_j^T(t)$ is the transverse component (direction perpendicular to $\mathbf{k}$) of the velocity of the jth particle at time t. From the Navier–Stokes equation (11) we find

$$[\partial J_t(k,t)/\partial t] = -\nu k^2 J_t(k,t) \quad (23)$$

where $\nu = \eta/\rho_0$. The corresponding frequency spectrum is a Lorentzian function.

$$J_t(k,\omega) = 2v_0^2\nu k^2\big/[\omega^2 + (\nu k^2)^2] \quad (24)$$

with $J(k, t=0) = v_0^2 = (\beta m)^{-1}$.
We see that at long wavelengths transverse current fluctuations in a fluid dissipate by simple diffusion at a rate given by νk^2.

IV. GENERALIZED HYDRODYNAMICS

The hydrodynamic description of fluctuations in fluids is expected to become inappropriate at finite values of (k,ω) when $kl \gtrsim 1$ and $\omega \gtrsim \omega_c$, where l is the collision mean free path and ω_c the mean collision frequency (see Section I). Nevertheless, we can extend the hydrodynamic description by allowing the thermodynamic coefficients in Eqs. (12) and (13) to become wavenumber dependent and the transport coefficients to become k- and ω-dependent. This method of extension is called generalized hydrodynamics.

The basic idea of generalized hydrodynamics can be simply presented by considering the case of the transverse current fluctuations. One of the fundamental differences between simple liquids and solids is that the former cannot support a shear stress, which is another way of saying that they have zero shear modulus. On the other hand, it is also known that at sufficiently short wavelengths or high frequencies shear waves can propagate through a simple liquid because then the system behaves like a viscoelastic medium. We have seen that according to hydrodynamics the frequency spectrum of the transverse current correlation function, (24), describes a diffusion process at all

frequencies. The absence of a propagating mode in (24) is an example of the inability of linearized hydrodynamics to treat viscoelastic behavior at finite (k, ω).

In the approach of generalized hydrodynamics we extend (23) by postulating the equation

$$\frac{\partial}{\partial t} J_t(k,t) = -k^2 \int_0^t dt' K_t(k, t-t') J_t(k,t') \qquad (25)$$

The kernel $K_t(k,t)$ is called a memory function; it is itself a time correlation function like $J_t(k,t)$. The role of K_t is to enable J_t to take on a short-time behavior that is distinctly different from its behavior at long times. It is reasonable that a quantity such as $K_t(k,t)$ should be present in the extension of hydrodynamics. With the introduction of a suitable $K_t(k,t)$, we expect that (25) will give shear wave propagation at finite (k,ω), while in the limit of small (k,ω) we recover Eq. (23).

On a phenomenological basis, without specifying completely $K_t(k,t)$ by a systematic derivation, we can require this function to satisfy requirements that incorporate certain properties of the function that we can readily derive. The two properties of $K_t(k,t)$ most relevant to the present discussion are

$$K_t(k, t=0) = (nm)^{-1} G_\infty(k) \qquad (26)$$

and

$$\lim_{k\to 0} \int_0^\infty dt\, K_t(k,t) = \nu \qquad (27)$$

where $G_\infty(k)$ is the high-frequency shear modulus. Both G_∞ and ν are actually properties of $J_t(k,\omega)$,

$$\frac{(kv_0)^2}{nm} G_\infty(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega\, \omega^2 J_t(k,\omega) \qquad (28)$$

$$2v_0^2 \nu = \lim_{\omega\to 0} \lim_{k\to 0} \left(\frac{\omega}{k}\right)^2 J_t(k,\omega) \qquad (29)$$

Moreover, Eq. (28) can be reduced to a kinetic contribution $(kv_0^2)^2$ and an integral over $g(r)$ and potential function derivative that can be evaluated by quadrature.

Equations (26) and (27) may be regarded as constraints or "boundary conditions" on $K_t(k,t)$, but by themselves they do not determine the memory function. Empirical forms have been proposed for $K_t(k,t)$ with adjustable parameters determined by imposing Eqs. (26) and (27). As an example, we consider the exponential or single relaxation time model,

$$K_t(k,t) = [G_\infty(k)/nm] \exp[-t/\tau(k)] \qquad (30)$$

where we are still free to specify the wavenumber-dependent relaxation time $\tau(k)$. Notice that Eq. (26) has already been incorporated. Applying Eq. (27) we obtain $\tau(k=0) = nm\nu/G_\infty(0)$, a quantity sometimes called the Maxwell relaxation time in viscoelastic theories. Furthermore, we expect $\tau(k)$ to be a decreasing function of k on the grounds that fluctuations at shorter wavelengths generally dissipate more rapidly. The simple interpolation expression

$$\frac{1}{\tau_t^2(k)} = \frac{1}{\tau_t^2(0)} + (kv_0)^2 \qquad (31)$$

would be consistent with this expectation and entails no further parameters. There exist more elaborate models for $\tau(k)$ as well as for $K_t(k,t)$, but the model Eq. (30) with (31) has the virtue of simplicity. Then Eq. (25) gives

$$J_t(k,\omega) = \frac{2v_0^2 k^2 K_t(k,0)}{\tau_t(k)} \times \left\{ \left[\omega^2 - \left\{ k^2 K_t(k,0) - \frac{1}{2\tau_t^2(k)} \right\} \right]^2 + \left[k^2 K_t(k,0) - \frac{1}{4\tau_t^2(k)} \right] \Big/ \tau_t^2(k) \right\}^{-1} \qquad (32)$$

The effect of the memory function now may be seen in the spectral behavior of $J_t(k,\omega)$. Whenever

$$k^2 K_t(k,0) > \frac{1}{2\tau_t^2(k)} \qquad (33)$$

there will exist a finite frequency, where the denominator in Eq. (32) is a minimum, and $J_t(k,\omega)$ will show a resonant peak. The resonant structure indicates a propagating mode associated with shear waves. Notice that Eq. (33) cannot hold at sufficiently small k; thus, in the long wavelength limit Eq. (32) can only describe diffusion, in agreement with Eq. (24). Figure 2 shows the data of molecular dynamics simulation; we see clear evidence of the onset of shear waves as k increases.

Generalized hydrodynamic descriptions for other time correlation functions also can be developed by using memory function equations such as Eq. (25). We will briefly summarize the results for density and longitudinal current fluctuations. The continuity equation, Eq. (6), is an exact expression, unlike the Navier–Stokes or the energy transport equation. One of its implications is a rigorous relation between the density correlation function, $F(k,t)$, and the longitudinal current correlation function $J_l(k,t)$. The latter is defined in a similar way as Eq. (22), with the transverse component v_j^{T} replaced by the longitudinal component (direction parallel to $\mathbf{k}$). In terms of the dynamic structure factor $S(k,\omega)$, the relation is

$$J_l(k,\omega) = (\omega/k)^2 S(k,\omega) \qquad (34)$$

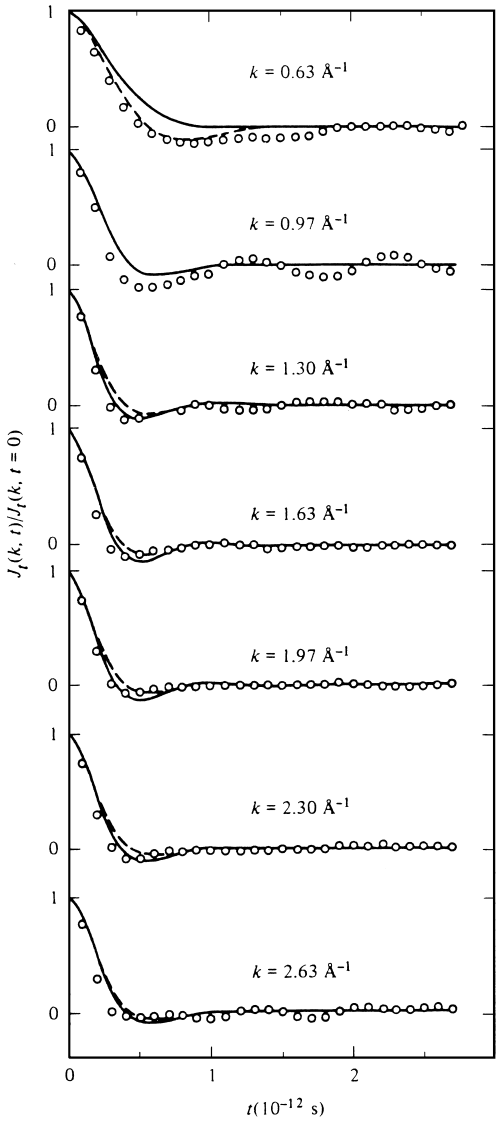


FIGURE 2 Normalized transverse current correlation function of liquid argon at various wavenumbers, molecular dynamics simulation data (circles), and exponential memory function model with $\tau_t(k)$ given by Eq. (41) (solid curves) and by a more elaborate expression (dashed curves).

Since this holds in general, we will focus our attention on $J_l(k, \omega)$. For purposes of illustration we assume that temperature fluctuations can be ignored. This means that we can set $T_1 = 0$ in Eq. (11) and obtain

$$J_l(k, \omega) = 2v_0^2 \frac{(\omega k)^2}{[\omega^2 - (c_T k)^2]^2 + [\omega \nu k^2]^2} \tag{35}$$

with $c_T = c_0/\nu$ being the isothermal sound speed. We see that in the hydrodynamic description the longitudinal current fluctuations, in contrast to the transverse current fluctuations, propagate at a frequency essentially given by $\omega \sim c_T k^2$. If temperature fluctuations were not neglected, the propagation frequency would be $c_0 k$ and the damping constant governed by the sound attenuation coefficient Γ [cf. Eq. (21)] instead of ν as in Eq. (35).

The inadequacy of the hydrodynamic description Eq. (35) at finite (k, ω) values is more subtle than is the case of $J_t(k, \omega)$. We find that Eq. (35) gives an overestimate of the damping of fluctuations, and it does not describe any of the effects associated with the intermolecular structure as manifested through the static structure factor $S(k)$. The extension of Eq. (35) can proceed if we write

$$\frac{\partial J_l(k, t)}{\partial t} = -\int_0^t dt' K_l(k, t - t') J_l(k, t') \tag{36}$$

with

$$K_l(k, t) = \frac{(k v_0)^2}{S(k)} + k^2 \phi_l(k, t) \tag{37}$$

The form of $K_l(k, t)$ is motivated by the coupling of Eqs. (10) and (11), and the generalization of the isothermal compressibility x_T, $n k_B T x_T \to S(k)$. Combining Eqs. [(35) and (36)] gives

$$J_l(k, \omega) = \frac{2v_0^2 (\omega k)^2 \phi_l'(k, \omega)}{\left[\omega^2 - \dfrac{(k v_0)^2}{S(k)} + \omega k^2 \phi_l''(k, \omega)\right]^2 + [\omega k^2 \phi_l'(k, \omega)]^2} \tag{38}$$

where ϕ_l'' and ϕ_l' are the real and imaginary parts of

$$\phi_l(k, s) = \int_0^\infty dt\ e^{-st} \phi_l(k, t) \tag{39}$$

with $s = i\omega$, and they describe the dissipative and reactive responses, respectively.

It is evident from a comparison of Eq. (38) with Eq. (35) that in addition to the generalization of the isothermal compressibility, the longitudinal viscosity has become a complex k- and w-dependent quantity. Through $\phi_l(k, t)$ we can again introduce physical models and use various properties to determine the k dependence. One way to characterize the breakdown of hydrodynamics in the case of $J_l(k, \omega)$ is to follow the frequency of the propagating mode as k increases. Notice first that by virtue of Eq. (34) $J_l(k, \omega)$ always shows a peak at a nonzero frequency. At small k this peak is associated with sound propagation. If we define

$$c(k) = \frac{\omega_m(k)}{k} \tag{40}$$

where $\omega_m(k)$ is the peak position, then $c(k)$ in the long wavelength limit is the adiabatic sound speed. This being the case, it is reasonable to regard Eq. (40) as the speed at which collective modes propagate in the fluid at any wavenumber. In terms of $c(k)$ we have a well-defined

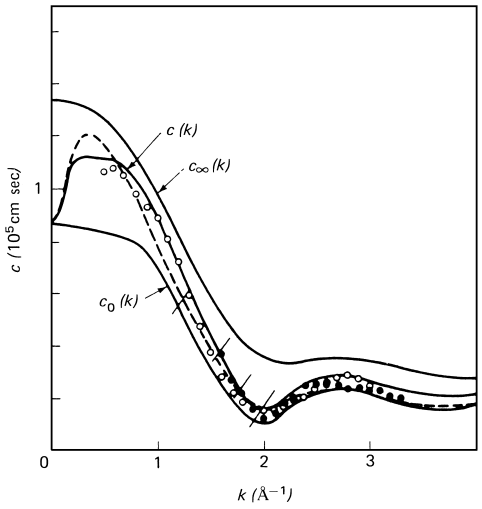


FIGURE 3 Variation of propagation velocities with wavenumber in liquid argon. Generalized hydrodynamics results are given as the solid curve denoted by $c(k)$ and by the dashed curve, neutron scattering measurements are denoted by the closed circles and slash marks, and computer simulation data are denoted by open circles. The quantities $c_0(k)$ and $c_\infty(k)$ are defined in the text.

quantity for discussing the variation of propagation speed at finite k. Notice that we do not refer to the propagating fluctuations at finite k as sound waves, because the latter are excitations that manifest clearly in $S(k, \omega)$.

There exist computer simulation results and neutron inelastic scattering data on simple liquids from which $c(k)$ can be determined. Figure 3 shows a comparison of these results with a generalized hydrodynamics calculation. Also shown are the adiabatic sound speed $c_0(k)$ and the high-frequency sound speed $c_\infty(k)$,

$$c_0(k) = v_0[\gamma/S(k)]^{1/2} \tag{41}$$

$$c_\infty(k) = \left\{\frac{1}{nm}\left[\frac{4}{3}G_\infty(k) + K_\infty(k)\right]\right\}^{1/2} \tag{42}$$

where K_∞ is the high-frequency bulk modulus. It is seen in Fig. 3 that $c_0(k)$ and $c_\infty(k)$ provide lower and upper bounds on $c(k)$. The fact that $c(k)$ deviates from both may be attributed to dynamical effects, which cannot be described through static properties such as in Eqs. (41) and (42). Relative to the adiabatic sound speed $c_0(k \to 0)$ we see in $c(k)$ first an enhancement as k increases up to about $1\ A^{-1}$, then a sharp decrease at larger k. The former behavior, a positive dispersion, is due to shear relaxation, whereas the latter, a strong negative dispersion, is due to structural correlation effects represented by $S(k)$. From this discussion we may conclude that an expression such as Eq. (38), with rather simple physical models for $\phi_l(k, t)$, provides a semiquantitatively correct description of density and current fluctuations at finite (k, ω).

V. KINETIC THEORY

In the theory of particle and radiation transport in fluids there exists a well established connection between the continuum approach as represented by the hydrodynamics equations and the molecular approach as represented by kinetic equations in phase space, an example of which is the Boltzmann equation in gas dynamics. Through this connection we can obtain expressions for calculating the input parameters in the continuum equations, such as the transport coefficients in Eqs. (11) and (12). We can also solve the kinetic equations directly to analyze thermal fluctuations at finite (k, ω), and in this way take into account, explicitly, the effects of spatial correlations and detailed dynamics of molecular collisions. In contrast to generalized hydrodynamics, the kinetic theory method allows us to derive, rather than postulate, the space–time memory functions like $K(k, t)$.

The essence of the kinetic theory description is that particle motions are followed in both configuration and momentum space. Analogous to Section II we begin with the phase space density

$$A(\mathbf{r}\mathbf{p}t) = \sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{R}_i(t))\delta(\mathbf{p} - \mathbf{P}_i(t)) \tag{43}$$

and the time-dependent phase-space density correlation function [cf. Eq. (7)]

$$C(\mathbf{r} - \mathbf{r}', \mathbf{p}\mathbf{p}', t) = \langle \delta A(\mathbf{r}\mathbf{p}t)\delta A(\mathbf{r}'\mathbf{p}'0)\rangle \tag{44}$$

with $\langle A\rangle = nf_0(p)$. The fundamental quantity in the analysis is now $C(\mathbf{r}, \mathbf{p}\mathbf{p}', t)$, from which the time correlation functions of Section II can be obtained by appropriate integration over the momentum variables. For example,

$$G(\mathbf{r}, t) = \int d^3p d^3p' C(\mathbf{r}, \mathbf{p}\mathbf{p}', t) \tag{45}$$

Various methods have been proposed to derive the equation governing $C(r, \mathbf{p}\mathbf{p}', t)$. All the results can be put into the generic form

$$\left(z - \frac{\mathbf{k}\cdot\mathbf{p}}{m}\right) C(k\mathbf{p}\mathbf{p}'z) - \int d^3p''\phi(k\mathbf{p}\mathbf{p}''z)C(k\mathbf{p}''\mathbf{p}'z) = -iC_0(k\mathbf{p}\mathbf{p}') \tag{46}$$

where

$$C(k\mathbf{p}\mathbf{p}'z) = \int d^3r \int_0^\infty dt\ e^{i(\mathbf{k}\cdot\mathbf{r} - zt)} C(\mathbf{r}, \mathbf{p}\mathbf{p}', t) \tag{47}$$

with the initial condition

$$C_0(k\mathbf{p}\mathbf{p}') = \int d^3r\, e^{i\mathbf{k}\cdot\mathbf{r}} C(\mathbf{r}, \mathbf{p}\mathbf{p}', t = 0) = nf_0(p)\delta(\mathbf{p} - \mathbf{p}') + n^2 f_0(p) f_0(p') h(k) \tag{48}$$

and $nh(k) = S(k) - 1$. In Eq. (46) the function $\phi(k\mathbf{pp}'z)$ is the phase-space memory function, which plays the same role as the memory function $K(k,t)$ in Eq. (25) or Eq. (36). It contains all the effects of molecular interactions. If ϕ were identically zero, then Eq. (46) would describe a noninteracting system in which the particles move in straight line trajectories at constant velocities. We can also think of ϕ as the collision kernel in a transport equation.

There are a number of formal properties of ϕ pertaining to symmetries, conservation laws, and asymptotic behavior, which one can analyze. Also, explicit calculations have been made under different conditions, such as low density, weak coupling, or relaxation time models. In general, it is useful to separate ϕ into an instantaneous, or static, part and a time-varying, or collisional, part,

$$\phi(k\mathbf{pp}'z) = \phi^{(s)}(k\mathbf{p}) + \phi^{(c)}(k\mathbf{pp}'z) \tag{49}$$

where

$$\phi^{(s)}(k\mathbf{p}) = -\frac{\mathbf{k}\cdot\mathbf{p}}{m} nf_0(p)C(k) \tag{50}$$

The quantity $C(k) = [S(k) - 1]/nS(k)$ is known as the direct correlation function. Physically $\phi^{(s)}$ represents the effects of mean field interactions with $nC(k)$ as the effective potential of the fluid system.

The calculation of $\phi^{(c)}$ is a difficult problem because we have to deal with the details of collision dynamics. It can be shown that in the limit of low densities, low frequencies, and small wavenumbers, $\phi^{(c)}$ reduces to the collision kernel in the linearized Boltzmann equation. This connection is significant because the Boltzmann equation is the fundamental equation in the study of transport coefficients and of the response of a gas to external perturbations.

The basic assumption underlying the Boltzmann equation is that intermolecular interactions can be treated as a sequence of uncorrelated binary collisions. This assumption renders the equation much more tractable, but it also limits the validity of the equation to low-density gases. Figure 4 shows the frequency spectrum of density fluctuations in xenon gas at 349.6 K and 1.03 atm calculated according to the procedure:

$$S(k,\omega) = \frac{1}{\pi} Re \int d^3p d^3p' C(k\mathbf{pp}'z)_{z=i\omega} \tag{51}$$

where C is determined from Eq. (46) with $\phi^{(c)}$ given by the binary collision kernel for hard sphere interactions. At such a low density it is valid to ignore $\phi^{(s)}$ and the second term in Eq. (48).

Also shown in Fig. 4 are the experimental data from light scattering spectroscopy. The good agreement is evidence that the linearized Boltzmann equation provides an accurate description of thermal fluctuations in low-density gases in the kinetic regime where $kl \sim 1$. The agreement is less satisfactory when the data are compared with the results of hydrodynamics; in this case the calculated spectrum shows essentially no structure. This again indicates that at finite (k,ω) the hydrodynamic theory overestimates the damping of density fluctuations. Generally speaking, kinetic theory calculations have been quantitatively useful in the analysis of light scattering experiments on gases and gas mixtures.

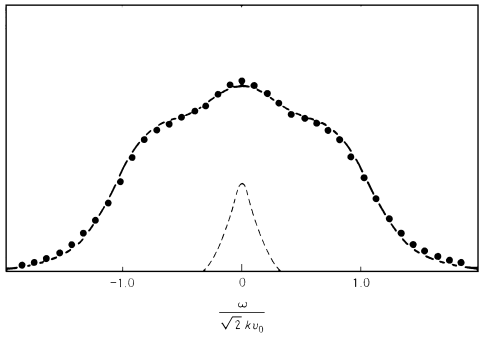


FIGURE 4 Frequency spectrum of dynamic structure factor in xenon gas at 349.6 K and 1.03 atm; light scattering data for 6328 Å incident light and scattering angle of 169.4° are shown as closed circles while the full curve denotes results obtained using the linearized Boltzmann equation for hard spheres. Calculated spectrum has been convolved with the resolution function shown by the dashed curve.

For moderately dense systems, typically fluids at around the critical density, the Boltzmann equation needs to be modified to take into account the local structure of the fluid. In the case of hard spheres, the modified equation generally adopted is the generalized Enskog equation, which involves $g(\sigma)$, the pair distribution function at contact (with σ the hard sphere diameter); the collision term differs from the collision integral in the linearized Boltzmann equation for hard spheres only in the presence of two phase factors, which represent the nonlocal spatial effects in collisions between molecules of finite size.

Figure 5 shows the frequency spectra of density fluctuations obtained from simulation and kinetic theory at rather long wavelengths in hard sphere fluids and at three densities, corresponding roughly to half the critical density, 1.7 times critical density, and liquid density at the triple point. The $k\sigma$ values are such that using the expression $l^{-1} = \sqrt{2}\pi n\sigma^2 g(\sigma)$ for the collision mean free path, we find that for the three cases (a)–(c) a molecule on the average would have suffered about 1, 5, and 20 collisions, respectively, in traversing a distance equal to the hard sphere diameter. On this basis we might expect the spectra in (b) and (c) to be dominated by hydrodynamic behavior, while that in (a) should show significant deviations.

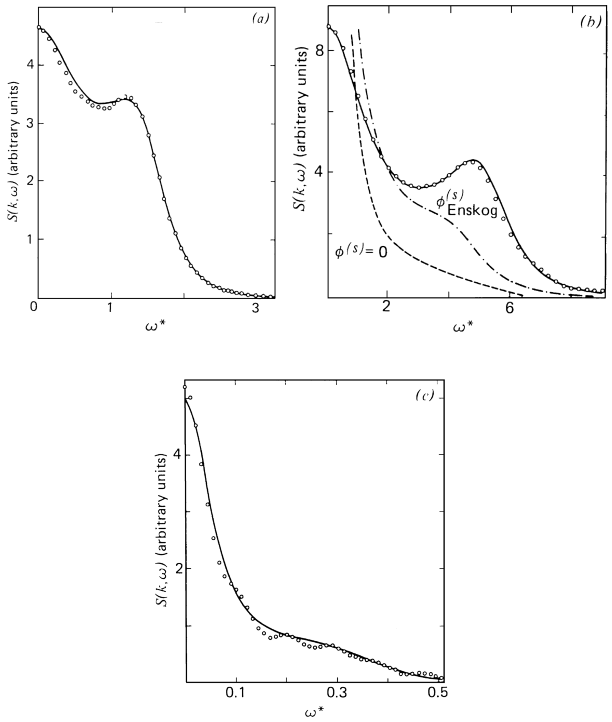


FIGURE 5 Frequency spectra of dynamic structure factor $S(k, \omega)$ in hard sphere fluids at three densities; simulation data are shown as open circles while the solid curves denote results obtained using the generalized Enskog equation. Dimensionless frequency ω^* is defined as $\omega\tau_E/k\sigma$, with the Enskog collision time $\tau_E^{-1} = 4\sqrt{\pi}n v_0 \sigma^2 g(\sigma)$. Only inputs to the calculations are $S(k)$ and $g(\sigma)$, which can be obtained from the simulation data. In (b) the effects of ignoring entirely the static part of the memory function and of using the conventional Enskog equation are also shown. For (a) $n\sigma^3 = 0.1414$, $k\sigma = 0.412$, $S(k) = 0.563$; $g(\sigma) = 1.22$; (b) $n\sigma^3 = 0.471$, $k\sigma = 0.616$; $S(k) = 0.149$, $g(\sigma) = 4.98$; and (c) $n\sigma^3 = 0.884$, $k\sigma = 0.759$, $S(k) = 0.0271$, $g(\sigma) = 2.06$.

The theoretical curves in Fig. 5 are kinetic model solutions to the generalized Enskog equation. They are seen to describe quantitatively the computer simulation data. We could have expected good agreement in the lowest density case, which is nevertheless two orders of magnitude higher in density than a gas under standard conditions. That the theory is still accurate at condition (b) is already somewhat unexpected. So it is rather surprising that a kinetic theory that treats the interactions as only uncorrelated binary collisions is applicable at liquid density, as shown in (c). Indeed, at three times the present value of $k\sigma$ a characteristic discrepancy appears in the high-density case, as shown in Fig. 6.

The failure of the generalized Enskog equation to account properly for the simulation results at low frequencies can be traced to the presence of a slower decaying component in the data for $F(k, t)$. It seems reasonable to associate this with the relaxation of clusters of particles, which should become important at high densities. Just like the onset of shear wave propagation, this characteristic feature is part of the viscoelastic behavior expected of dense fluids. In order to describe such effects in the present context, it is now recognized that correlated collisions will have to be included in the kinetic equation. Aside from density and thermal fluctuations, it is also known that the transport coefficients derived from the Enskog equation are in error up to a factor of 2 at the liquid density when compared to computer simulation data hard spheres. Moreover, simulation studies have revealed a nonexponential, long-time decay of the velocity autocorrelation function that cannot be explained by the Enskog theory.

Any attempt to treat correlated collision effects necessarily leads to nonlinear kinetic equations. For practical calculations it appears that only the correlated binary collisions, called ring collisions, are tractable. To incorporate these dynamical processes in the kinetic theory, we can develop a formalism wherein $\phi^{(c)}$ is given as the sum $\phi^{(c)} = \phi_E + \phi_R$, where ϕ_E is the memory function

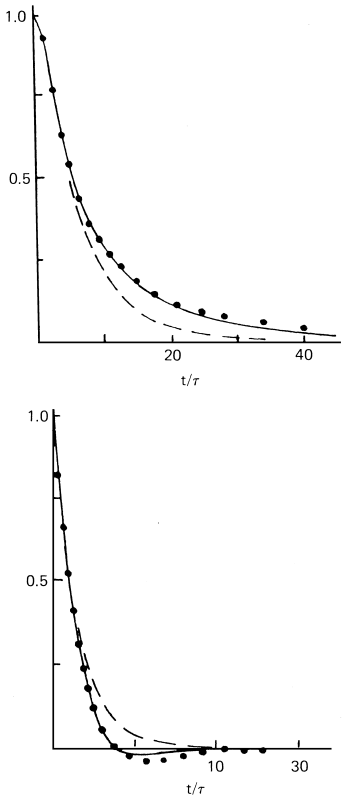


FIGURE 6 The density correlation function and normalized transverse current correlation in a hard sphere fluid at a density of $n\sigma^3 = 0.884$. The k value is 2.28 σ^{-1}. Computer simulation data are given by the circles, while calculations using the generalized Enskog equation or the mode coupling theory are denoted by the dashed and solid curves, respectively.

for the generalized Enskog equation, and ϕ_R describes the ring collision contribution. In essence ϕ_R can be expressed schematically as $\phi_R = VCCV$, where V is an effective interaction, which involves the actual intermolecular potential and the equilibrium distribution function of the fluid, and C is the phase space correlation function. The important point to note is that the memory function now depends quadratically on C, thereby making Eq. (46) a nonlinear kinetic equation. The appearance of nonlinearity, or feedback effects, is not so surprising when we recognize that in a dense medium the motions of a molecule will have considerable effects on its surroundings, which in turn will react and influence its subsequent motions. The inclusion of correlated collisions is a significant development in the study of transport called renormalized kinetic theory.

The presence of ring collisions unavoidably makes the analysis of time correlation functions considerably more difficult. Nevertheless, it can be shown analytically that we obtain a number of nontrivial collective properties characteristic of a dense fluid, such as a power law decay of the velocity autocorrelation function, and nonanalytic density expansions of sound dispersion and transport coefficients.

VI. MODE COUPLING THEORY

There exists another method of analyzing time correlation functions, which has common features with both generalized hydrodynamics and renormalized kinetic theory. In this approach we formulate an approximate expression for the space–time memory function that is itself nonlinear in the time correlation functions. The method is called mode coupling because the correlation functions describe the hydrodynamic modes, the conserved variables of density, momentum, and energy, in the small (k, ω) limit, and they are brought together to represent higher order correlations, which are important in a strongly coupled system such as a liquid.

The mode coupling approach has been particularly successful in describing the dynamics of dense fluids; it is the only tractable microscopic theory of dense simple fluids. To describe the mode coupling formalism, we consider the density correlation function or its Laplace transform

$$S(k, z) \equiv i \int_0^\infty dt\ e^{izt} F(k, t) \equiv \mathscr{L}[F(k, t)] \qquad (52)$$

and similarly for $J_l(k, z)$, the longitudinal current correlation function. Using the continuity equation we find

$$S(k, z) = -\frac{S(k)}{z} + \left(\frac{k}{z}\right)^2 J_l(k, z) \qquad (53)$$

which is just another form of Eq. (34). Now we write an equation for $J_l(k, z)$ [cf. Eq. (36)] in the form

$$J_l(k, z) = -v_0^2 \left[z - \frac{\Omega_0^2(k)}{z} + D(k, z) \right]^{-1} \qquad (54)$$

where $\Omega_0^2(k) = (kv_0)^2/S(k)$ and $D(k, z)$ is the memory function. Combining Eqs. (53) and (54) gives

$$S(k, z) = -S(k) \left[z - \frac{\Omega_0^2(k)}{z + D(k, z)} \right]^{-1} \qquad (55)$$

which is an exact equation. The basic assumption underlying the mode coupling theory is the approximate expression derived for $D(k, z)$. In essence one obtains

$$D(k, z) \simeq \frac{\Omega_0^2(k)}{v} + \int d^3k'\, V(k, k') \mathscr{L}[F(k', t) \times F(|\mathbf{k} - \mathbf{k}'|, t)] \qquad (56)$$

where v is a characteristic collision frequency usually taken from the Enskog theory, and $V(k, k')$ is an effective interaction. Equation (56) is an example of a two-mode coupling approximation involving two density modes $F(k, t)$. Depending on the problem, we can have other products containing modes from the group $\{F(k, t), F_s(k, t), J_t(k, t), J_l(k, t)$, and the energy fluctuation$\}$, and for each mode coupling term there will be an appropriate vertex interaction $V(k, k')$.

The calculation of $S(k, z)$ is fully specified by combining Eqs. (55) and (56). By expressing the memory function back in terms of the correlation function, we obtain a self-consistent description capable of treating feedback effects. These are the effects that become important at high densities, and that we have tried to treat in the kinetic theory approach through the ring collisions.

Mode coupling calculations were first applied to analyze the transverse and longitudinal current correlation functions in liquid argon and liquid rubidium. The theory was found to give a satisfactory account of the computer simulation results on shear wave propagation in J_t and the dispersion behavior of $\omega_m(k)$ in J_l. The theory was then reformulated for the case of hard spheres and extensive numerical results were obtained and compared in detail with simulation data. It was shown that the viscoelastic behavior discussed previously, which could not be explained by the generalized Enskog equation, is now well described. The improvement due to mode coupling can be seen in Fig. 6.

Another problem where the capability of mode coupling analysis to treat dense medium effects can be demonstrated is the Lorentz model. This is the study of the diffusion of a tagged particle in a random medium of stationary scatterers and of its localization when the scatterer density n exceeds a critical value n_c. The system can be

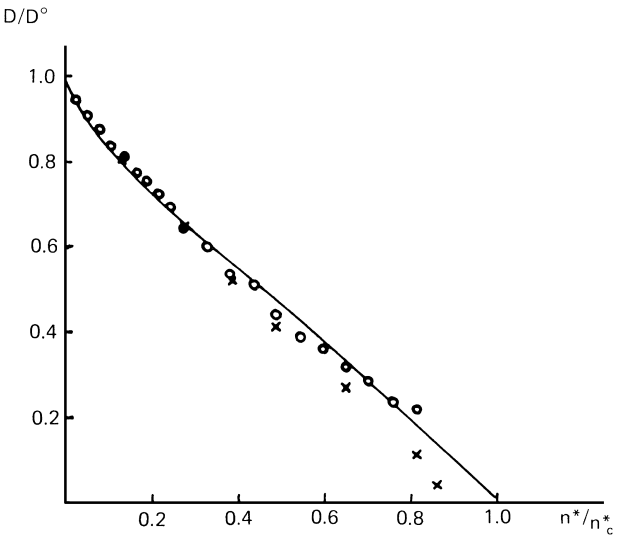


FIGURE 7 Density variation of the diffusion coefficient in the two-dimensional Lorentz model where the hard disks can overlap: mode coupling theory (solid curve) and computer simulation data. D^0 is the diffusion coefficient given by the Enskog theory.

characterized by the diffusion coefficient of the tagged particle D, which plays the role of an order parameter. The model then exhibits two distinct phases, a "diffusion" phase $D \neq 0$, when $n < n_c$, and for $n > n_c$ a "localization" phase with $D = 0$.

Figure 7 shows the density variation of D in the case of the two-dimensional Lorentz model with hard disk scatterers that can overlap. The mode coupling theory gives satisfactory results if the density is scaled according to n_c. As for the prediction of n_c, the theory gives $n_c^* = n\sigma^d = 0.64$ and 0.72 for $d = 2$ and $d = 3$, respectively, while molecular dynamics simulations give 0.37 and 0.72. Here the tagged particle and the stationary scatterers are both hard spheres of diameter σ, and d is the dimensionality of the system. The fact that the theory does not give an accurate value for n_c in two dimensions indicates that the statistical distribution of system configurations in which the particle becomes trapped requires a more complicated treatment than the simplest mode coupling approximation. On the other hand, the density variation of the velocity autocorrelation function observed by simulation, particularly its nonexponential decay at long times for $n < n_c$, can be calculated very satisfactorily.

In view of the successful attempts at describing the dense, hard sphere fluids and the Lorentz model, we might wonder what mode coupling theory will give at densities beyond the normal liquid density, typically taken to be the triple point density of a van der Waals liquid, $n^* = n\sigma^3 = 0.884$. On intuition alone we expect that as the atoms in the fluid are pushed more closely against each other, structural rearrangement becomes more and more difficult so that at a certain density the local structure will no longer relax on the time scale of observation. This condition of structural arrest is a fundamental characteristic of solidification, and it is appropriate to ask if mode coupling theory can describe such a highly cooperative process. Indeed, a certain self-consistent approximation in mode-coupling theory will lead to a model that exhibits a freezing transition. The signature of the transition is that the system becomes nonergodic at a critical value of the density or temperature.

To demonstrate that the mode-coupling formalism can describe a transition from ergodic to nonergodic behavior, we consider a schematic model for the normalized dynamic structure factor $\varphi(z) \equiv S(k,z)/S(k)$, and ignore the wavenumber dependence in the problem. In analogy with Eq. (55) we write

$$-\varphi(z)^{-1} = z + K(z) \tag{57}$$

$$-\Omega_0^2 K(z)^{-1} = z + M(z) \tag{58}$$

with $M(z)$ playing the role of $D(k,z)$. We will consider two different approximations to the memory function $M(z)$,

$$M(z) \simeq M_0(z) + m(z) \tag{59}$$

and

$$M(z) \simeq M_0(z) + \frac{m(z)}{1 - \Delta(z)\, m(z)} \tag{60}$$

with

$$M_0(t) = \bar{\omega}\, \delta(t) \tag{61}$$

$$m(t) = 4\lambda\Omega_0^2\, F^2(t) \tag{62}$$

$$\Delta(t) = \lambda' F(t)\, J(t) \tag{63}$$

Equations (59) and (62) constitute the original mode-coupling approximation, henceforth denoted as the LBGS model, in which only the coupling of density fluctuation modes, with $F(t)$ defined by Eq. (18), is considered. Equations (60) and (63) constitute an extension in which the coupling to longitudinal current modes, with $J(t)$ given by Eq. (34), is also considered. We will refer to this as the extended mode-coupling approximation. In both models $M_0(z) = i\bar{\omega}$ is the Enskog-theory contribution to the memory function, where $\bar{\omega}$ is an effective collision frequency. The coupling coefficients λ and λ' will be treated as density- and temperature-dependent constants. Comparing the two models, we see that the difference lies in the presence of $\Delta(z)$ in Eq. (60).

It may seem remarkable that an apparently simple approximation of coupling two density modes can provide a dynamical model of freezing. To see how this comes about, notice that the quantity of interest in the analysis is the relaxation behavior of the time-dependent

density correlation function $G(r, t)$, or its Fourier transform $F(t) = F(k, t)$. Under normal conditions one expects $F(t \to \infty) = 0$ because all thermal fluctuations in an equilibrium system should die out if one waits long enough. When freezing occurs, this condition no longer holds as some correlations now can persist for all times. The condition that $F(t)$ stays finite as $t \to \infty$ means the system has become nonergodic. To see that Eq. (59) can give such a transition, we look for a solution to the closed set of equations (57), (58), and (59) of the form

$$\varphi(z) = -f/z + (1 - f)\varphi_v(z) \tag{64}$$

where the first term is that component of $F(t)$ which does not vanish at long times, $F(t \to \infty) = f$, and $\varphi_v(z)$ is a well-behaved function that is not singular at small z. Since $\varphi_v(z)$ is not pertinent to our discussion, we do not need to show it explicitly. Inserting this result into (59) yields

$$Mz = -4\lambda\Omega_0^2 f/z + M_v(z) \tag{65}$$

with $M_v(z)$ representing all the terms that are nonsingular at small z; we obtain

$$\varphi(z) = -\frac{4\lambda f^2}{1 + 4\lambda f^2}\frac{1}{z} - \Lambda(z)\frac{1}{1 + 4\lambda f^2} \tag{66}$$

with $\Lambda(z)$ also well behaved. For Eqs. (64) and (66) to be compatible, we must require

$$f = \frac{4\lambda f^2}{1 + 4\lambda f^2} \tag{67}$$

This is a simple quadratic equation for f, with solution $f = 1/2 + (1/2)(1 - 1/\lambda)^{1/2}$. Therefore, we see that in order for the postulated form of the density correlation function to be acceptable solution to Eqs. (57), (58), and (59), f must be real, or $\lambda > 1$.

The implication of this analysis is that in the LBGS model the ergodic phase is defined by the region $\lambda < 1$, where the nondecaying component f must vanish, and a nonergodic phase exists for $\lambda > 1$. The onset of nonergodicity signifies the freezing in of some of the structural degrees of freedom in the fluid; therefore, it may be regarded as a transition from a liquid to a glass. The origin of this transition is purely dynamical since it arises from a nonlinear feedback mechanism introduced through $m(z)$. The freezing or localization of the particles shows up as a simple pole in the low-frequency behavior of $\varphi(z)$, a consequence of the fact that $M(z) \approx 1/z$ at low frequencies.

The LBGS model is the first mode-coupling approximation providing a dynamical description of an ergodic to nonergodic transition. The transition also has been derived using a nonlinear fluctuating hydrodynamics formulation. Analysis of the LBGS model shows that the diffusion coefficient D has a power-law density dependence, $D \sim (n_c - n)^{\alpha}$, with exponent $\alpha \simeq 1.76$, and correspondingly the reciprocal of the shear viscosity coefficient η behaves in the same way. There exist experimental and molecular dynamics simulation data that provide evidence supporting the density and temperature variation of transport coefficients predicted by the model. Specifically, diffusivity data for the supercooled liquid methy-cyclohexane and for hard-sphere and Lennard–Jones fluids obtained by simulation are found to have density dependence that can be fitted to the predicted power law. The fact that the mode-coupling approximation is able to give a reasonable description of transport properties in liquids at high densities and low temperatures beyond the triple point is considered rather remarkable.

The LBGS model also has been found to provide the theoretical basis for interpreting recent neutron and light scattering measurements on dynamical relaxations in dense fluids. These experiments show that the temporal relaxation of the density correlation function $F(k, t)$ is nonexponential, $F(k, t) \simeq \exp[-t/\tau)^{\beta}]$, with β distinctly less than unity. This behavior of scaling, in the sense of F being a function of t/τ, where τ is a temperature-dependent relaxation time, and of stretching, in the sense of $\beta < 1$, is also given by Eq. (59) provided a term $\lambda'' F(t)$ is added to Eq. (62). Thus, the ability of the mode-coupling approximation to describe the dynamical features of relaxation in dense fluids has considerable current experimental support.

The successes of the approximation Eq. (59) notwithstanding, it does have an important physical shortcoming, namely, it does not treat the hopping motions of atoms when they are trapped in positions of local potential minima. These motions are expected to be dominant at sufficiently low temperature of supercooling; their presence means that the system should remain in the ergodic phase, albeit the relaxation times can become exceedingly long. For this reason, the predicted transition of the LBGS model is called the ideal glass transition; in reality one does not expect such a transition to be observed. The extended mode-coupling model, Eq. (60), in fact provides a cutoff for the ideal glass transition by virtue of the presence of $\Delta(z)$. One can see this quite simply from the small-z behavior of Eqs. (57), (58), and (60). With Δ nonzero, $\varphi(z)$ no longer has a singular component varying like $1/z$, so $F(t)$ will always vanish at sufficiently long times.

Even though the two approximations, Eqs. (59) and (60), give different predictions for the transition, one has to resort to numerical results in order to see the differences between the two models in their descriptions of $F(k, t)$ in the time region accessible to computer simulation and neutron and light scattering measurements. In Fig. 8 we show the intermediate scattering function $F(k, t)$ of a fluid calculated by simulation using a truncated Lennard–Jones interaction at various fluid densities

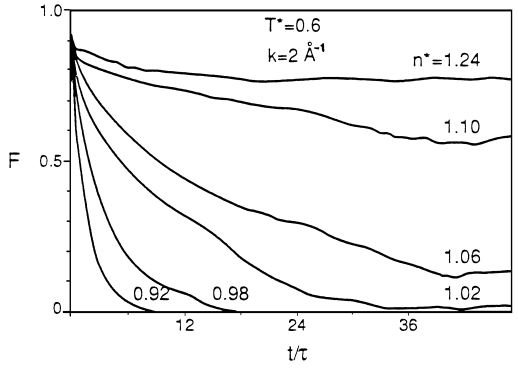


FIGURE 8 Relaxation of density correlation function $F(t)$ at wavenumber $k = 2\ A^{-1}$ obtained by molecular dynamics simulation at various reduced densities n^* and reduced temperature $T^* = 0.6$. Time unit τ is defined as $(m\sigma^2/\varepsilon)^{1/2}$.

($n^* = n\sigma^3$). One sees that as n^* increases, the relaxation of $F(k, t)$ becomes increasingly slow. Compared to these simulation results, the corresponding mode-coupling calculations, using a model equivalent to Eq. (59), show the same qualitative behavior of slowing down of relaxation; however, the mode-coupling model predicts a freezing effect that is too strong. This discrepancy is not seen in a model equivalent to Eq. (60). Thus, there is numerical evidence that the cutoff mechanism of the transition, represented by Δ, is rather significant.

To what extent can the dynamical features of supercooled liquids be described by mode-coupling models such as Eqs. (59) and (60)? Although these approximations seem to give semiquantitative results when compared to the available experimental and simulation results, it is also recognized that hopping motions should be incorporated in order that the theory be able to give a realistic account of the liquid-to-glass transition.

VII. LATTICE GAS HYDRODYNAMICS

Fluctuations extend continuously from the molecular level to the hydrodynamic scale, but we have seen that there are experimental and theoretical limitations to the ranges where they can be probed and computed. Indeed, no theory provides a fully explicit analytical description of space–time dynamics establishing the bridge between kinetic theory and hydrodynamic theory, and scattering techniques have limited ranges of wavelengths over which fluctuation correlations can be probed. With numerical computational methods one can realize molecular dynamics simulations that in principle, could cover the whole desired range, but in practice there are computation time and memory requirement limitations.

Lattice gas automata (LGA) are discrete models constructed as an extremely simplified version of a many-particle system where pointlike particles residing on a regular lattice move from node to node and undergo collisions when their trajectories meet at the same node. The remarkable fact is that, if the collisions occur according to some simple logical rules and if the lattice has the proper symmetry, this automaton shows global behavior very similar to that of real fluids. Furthermore, the lattice gas automaton exhibits two important features: (i) It usually resides on large lattices, and so possesses a large number of degrees of freedom; and (ii) its microscopic Boolean nature, combined with the (generally) stochastic rules that govern its microscopic dynamics, results in intrinsic fluctuations. Therefore, the lattice gas can be considered as a "reservoir of thermal excitations" in much the same way as an actual fluid, and so can be used as a "virtual laboratory" for the analysis of fluctuations, starting from a microscopic description.

A lattice gas automaton consists of a set of particles moving on a regular d-dimensional lattice $\mathcal{L}$ at discrete time steps, $t = n\Delta t$, with n an integer. The lattice is composed of V nodes labeled by the d-dimensional position vectors $\mathbf{r} \in \mathcal{L}$. Associated to each node there are b channels (labeled by indices $i, j, \ldots$, running from 1 to b). At a given time, t, channels are either empty (the occupation variable $n_i(\mathbf{r}, t) = 0$) or occupied by *one* particle [$n_i(\mathbf{r}, t) = 1$]. If channel i at node $\mathbf{r}$ is occupied, then there is a particle at the specified node $\mathbf{r}$, with a velocity $\mathbf{c}_i$. The set of allowed velocities is such that the condition $\mathbf{r} + \mathbf{c}_i \Delta t \in \mathcal{L}$ is fulfilled. It may be required that the set $\{\mathbf{c}_i\}_{i=1}^b$ be invariant under a certain group of symmetry operations in order to ensure that the transformation properties of the tensorial objects that appear in the dynamical equations are the same as those in a continuum [such as the Navier–Stokes equation (11)]. The "exclusion principle" requirement that the maximum occupation be of one particle per channel allows for a representation of the automaton configuration in terms of a set of *bits* $\{n_i(\mathbf{r}, t); i = 1, \ldots, b; \mathbf{r} \in \mathcal{L}\}$. The evolution rules are thus simply logical operations over sets of bits, which can be implemented in an exact manner in a computer.

The time evolution of the automaton takes place in two stages: propagation and collision. We reserve the notation $\mathbf{n}(\mathbf{r}, t) \equiv \{n_i(\mathbf{r}, t)\}_{i=1}^b$ for the precollisional configuration of node $\mathbf{r}$ at time t, and $\mathbf{n}^*(\mathbf{r}, t) \equiv \{n_i^*(\mathbf{r}, t)\}_{i=1}^b$ for the configuration after collision. In the propagation step, particles are moved according to their velocity

$$n_i(\mathbf{r} + \mathbf{c}_i\ \Delta t, t + \Delta t) = n_i^*(\mathbf{r}, t) \tag{68}$$

The (local) collision step is implemented by redistributing the particles occupying a given node $\mathbf{r}$ among the channels associated to that node, according to a given

prescription, which can be stochastic. The collision step can be represented symbolically by

$$n_i^*(\mathbf{r}, t) = \sum_{\sigma} \sigma_i \, \xi_{\mathbf{n}(\mathbf{r},t)\to\sigma} \tag{69}$$

where $\xi_{\mathbf{n}(\mathbf{r},t)\to\sigma}$ is a random variable equal to 1 if, starting from configuration $\mathbf{n}(\mathbf{r}, t)$, configuration $\boldsymbol{\sigma} \equiv \{\sigma_i\}_{i=1}^b$ is the outcome of the collision, and 0 otherwise. The physics of the problem is reflected in the choice of transition matrix $\xi_{\mathbf{s}\to\boldsymbol{\sigma}}$. Taking an average over the random variable (assuming homogeneity of the stochastic process in both space and time), and using Eq. (68), we obtain

$$n_i(\mathbf{r} + \mathbf{c}_i, t + 1) = \sum_{\boldsymbol{\sigma},\mathbf{s}} \sigma_i \langle \xi \rangle_{\mathbf{s}\to\boldsymbol{\sigma}} \delta[\mathbf{n}(\mathbf{r}, t), \mathbf{s}] \tag{70}$$

where automaton units ($\Delta t = 1$) are used. These microdynamic equations constitute the basis for the theoretical description of correlations in lattice gas automata.

Starting from Eq. (70), by performing an ensemble average over an arbitrary distribution of initial occupation numbers (denoted by angular brackets), one derives a hierarchy of coupled equations for the n-particle distribution functions, analogous to the BBGKY hierarchy in continuous systems. The first two equations in this hierarchy are

$$f_i(\mathbf{r} + \mathbf{c}_i, t + 1) = \sum_{\boldsymbol{\sigma},\mathbf{s}} \sigma_i \langle \xi \rangle_{\mathbf{s}\to\boldsymbol{\sigma}} \langle \delta(\mathbf{n}(\mathbf{r}, t), \mathbf{s}) \rangle, \tag{71}$$

$$\begin{aligned} f_{ij}^{(2)}&(\mathbf{r} + \mathbf{c}_i, \mathbf{r}' + \mathbf{c}_j, t + 1) \\ &= (1 - \delta(\mathbf{r}, \mathbf{r}')) \Bigg\{ \sum_{\boldsymbol{\sigma},\mathbf{s},\boldsymbol{\sigma}',\mathbf{s}'} \sigma_i \sigma_j' \langle \xi \rangle_{\mathbf{s}\to\boldsymbol{\sigma}} \langle \xi \rangle_{\mathbf{s}'\to\boldsymbol{\sigma}'} \\ &\quad \times \langle \delta(\mathbf{n}(\mathbf{r}, t), \mathbf{s}) \delta(\mathbf{n}(\mathbf{r}', t), \mathbf{s}') \rangle \Bigg\} \\ &\quad + \delta(\mathbf{r}, \mathbf{r}') \Bigg\{ \sum_{\mathbf{s},\boldsymbol{\sigma}} \sigma_i \sigma_j \langle \xi \rangle_{\mathbf{s}\to\boldsymbol{\sigma}} \langle \delta(\mathbf{n}(\mathbf{r}, t), \mathbf{s}) \rangle \Bigg\} \end{aligned} \tag{72}$$

where

$$f_i(\mathbf{r}, t) = \langle n_i(\mathbf{r}, t) \rangle \tag{73}$$

$$f_{ij}^{(2)}(\mathbf{r}, \mathbf{r}', t) = \langle n_i(\mathbf{r}, t) n_j(\mathbf{r}', t) \rangle \tag{74}$$

are the one- and two-particle distribution functions, respectively. The fluctuations of the channel occupation number are $\delta n_i(\mathbf{r}, t) = n_i(\mathbf{r}, t) - f_i(\mathbf{r}, t)$, and the corresponding pair correlation function reads

$$G_{ij}(\mathbf{r}, \mathbf{r}', t) = \langle \delta n_i(\mathbf{r}, t) \delta n_j(\mathbf{r}', t) \rangle \tag{75}$$

Using a cluster expansion and neglecting three-point correlations, the hierarchy of equations can be approximately truncated to yield the *generalized Boltzmann equation* for the single particle distribution function $f_i(\mathbf{r}, t)$,

$$\begin{aligned} f_i(\mathbf{r} + \mathbf{c}_i, t + 1) - f_i(\mathbf{r}, t) &= \Omega_i^{(1,0)}(\mathbf{r}, t) \\ &\quad + \sum_{k<1} \Omega_{i,kl}^{(1,2)}(\mathbf{r}, t) G_{kl}(\mathbf{r}, \mathbf{r}, t) \end{aligned} \tag{76}$$

and the *ring kinetic equation* for the equal-time pair correlation function

$$\begin{aligned} &G_{ij}(\mathbf{r} + \mathbf{c}_i, \mathbf{r}' + \mathbf{c}_j, t + 1) \\ &\quad - \sum_{kl} W_{ij,kl}(\mathbf{r}, \mathbf{r}', t) \, G_{kl}(\mathbf{r}, \mathbf{r}', t) \\ &\quad = B_{ij}(\mathbf{r}, t) \delta(\mathbf{r}, \mathbf{r}') \end{aligned} \tag{77}$$

On the right-hand side of (76) the first term $\Omega_i^{(1,0)}$ represents the discrete nonlinear Boltzmann collision term, and the second term contains a set of correlated collision sequences (*ring events*). In the ring equation (77), $W_{ij,kl}(\mathbf{r}, \mathbf{r}', t)$ is a product of two homogeneous propagators,

$$W_{ij,kl}(\mathbf{r}, \mathbf{r}', t) = \left(\delta_{ik} + \Omega_{i,k}^{(1,1)}(\mathbf{r}, t)\right)\left(\delta_{jl} + \Omega_{j,l}^{(1,1)}(\mathbf{r}', t)\right) \tag{78}$$

where $\Omega_{i,j}^{(1,1)} = \partial \Omega_i^{(1,0)} / \partial f_j$ is the linearized collision operator, and the on-node source term $B_{ij}(\mathbf{r}, t)$ is a function of $f_i(\mathbf{r}, t)$ and of $G_{ij}(\mathbf{r}, \mathbf{r}', t)$.

In order to have a full description of the dynamics of a fluid, temperature should be associated to the lattice gas, which is only possible if the model system possesses a velocity distribution. A minimal two-dimensional thermal LGA can be constructed in the following manner: Particles reside on the two-dimensional triangular lattice (with hexagonal symmetry), have unit mass, and undergo displacements with velocity moduli 1, $\sqrt{3}$, and 2 (in lattice unit length per time step), and so have energies $1/2$, $3/2$, and 2, respectively; particles at rest have zero energy. Speeds 1 and 2 correspond to displacements by one and two lattice unit lengths, respectively, in one time step along any of the six lattice directions, and speed $\sqrt{3}$ corresponds to displacements to the next nearest neighboring nodes along any of the six directions bisecting the lattice directions. This defines a lattice gas with basic conservation laws: mass, momentum, and energy, and correct symmetry.

The existence of spontaneous thermal fluctuations in this LGA is convincingly evidenced by the analysis of the dynamic structure factor $S(\mathbf{k}, \omega)$, which in the linearized Boltzmann approximation is given by

$$S(\mathbf{k}, \omega) = 2\Re \sum_{ij} \left(\frac{1}{e^{i\omega + 1\mathbf{k}\cdot\mathbf{c}} - 1 - \Omega} \right)_{ij} \kappa_j + \rho S(\mathbf{k}) \tag{79}$$

where Ω is the linearized Boltzmann collision operator, $\Re$ denotes the real part, ρ is the particle density per node,

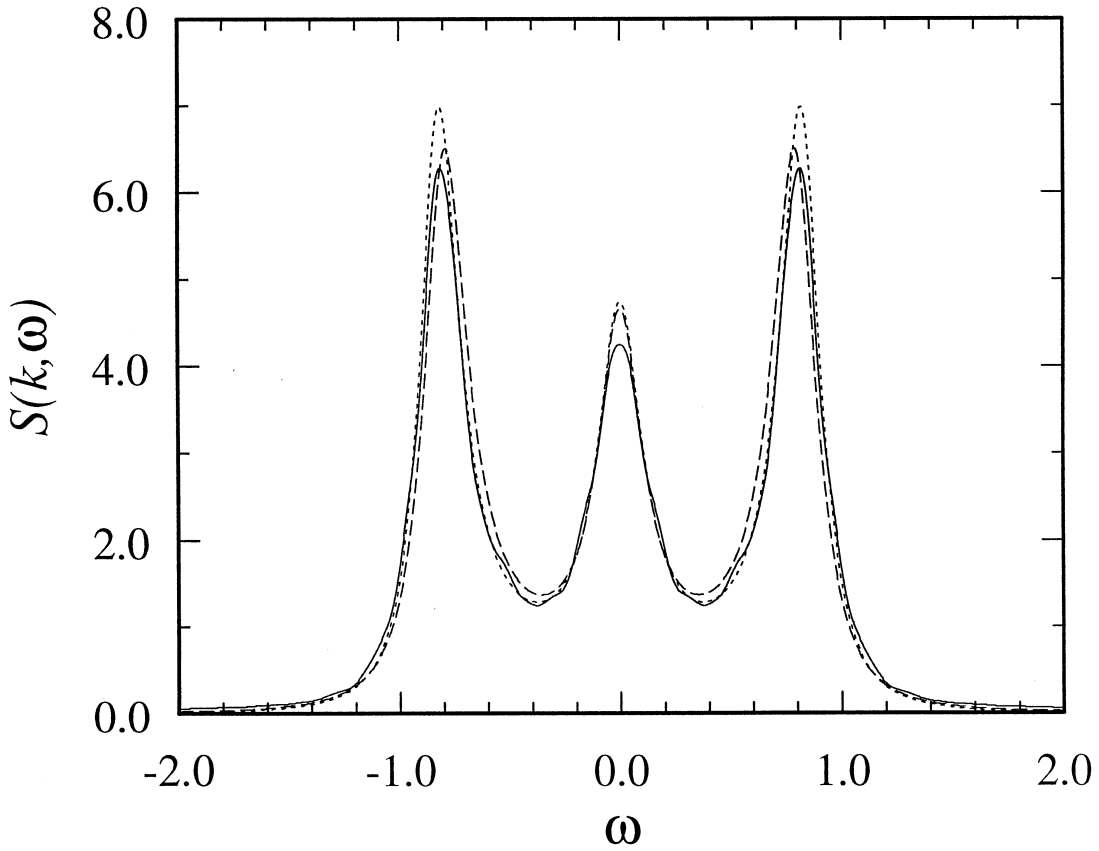


FIGURE 9 The lattice gas dynamic structure factor in the hydrodynamic regime: $S(k, \omega)$ as a function of the frequency ω at high density and low k. Comparison between the simulation results (solid line) and the theoretical predictions: hydrodynamic spectrum (dashed curve), and Boltzmann theory (dotted curve).

and $\kappa_j = f_j(1 - f_j)$, with f_j the average particle density per channel. The expression for $S(\mathbf{k}, \omega)$ is in general too complicated to be calculated analytically, but it can be used for direct numerical evaluation at all values of k. However, for small k values, that is, in the hydrodynamic regime where spatial and temporal variations are smooth, $S(\mathbf{k}, \omega)$ can be computed explicitly by low k expansion; the dynamic structure factor then takes the Landau–Placzek form as given by Eq. (21) (plus cross terms). A typical spectrum is given in Fig. 9. It shows that the spectral density of the lattice gas density fluctuations in the hydrodynamic domain exhibits the characteristic lineshape of the Rayleigh–Brillouin spectrum observed experimentally in real fluids.

A very intersting aspect of the lattice gas approach is that the eigenvalues $z_\mu(k)$ of the kinetic propagator

$$e^{-1\mathbf{k}\cdot\mathbf{c}}(1 + \Omega)\psi_\mu(\mathbf{k}) = e^{z_\mu(\mathbf{k})}\psi_\mu(\mathbf{k}) \qquad (80)$$

can be computed analytically for low k values, yielding the expressions for the transport coefficients, and numerically for any value of k. So here is a model system for which *all* modes can be computed as functions of k delineating the domains of validity of hydrodynamics, generalized hydrodynamics, and kinetic regime. For instance, it is found that the predictions of the lattice Boltzmann equation in which no small k- and/or ω-approximations have been made are valid over quite a wide range of wavenumbers, in good agreement with simulation results.

Lattice gas automata, as described so far, evolve according to an iterated sequence of mass- and momentum-preserved *local* collisions followed by propagation. *Nonlocal* interactions can be incorporated in the LGA dynamics via long-distance momentum transfer simulating attraction and/or repulsion between particles, by modifying the orientation of the velocity vectors from a diverging configuration to a converging configuration to simulate attractive forces and vice versa for repulsive forces. While in local collisions momentum redistribution is a node-located process, in nonlocal interactions (NLI) momentum is exchanged between two particles residing on nodes separated by a (fixed or variable) distance r: Mass is conserved locally, momentum is conserved globally. From the statistical mechanical viewpoint, LGAs with NLIs form an interesting class of models in that they include an

elementary process that is essential for "nonideal" behavior. At the macroscopic level, the main feature exhibited by LGAs with NLIs is a "liquid–gas"-type phase separation with bubble and drop formation.

The dynamics of LGA virtual particles is not governed by Newton's equation of motion, and the concepts of force and potential cannot be used in the sense of classical mechanics. Moreover, in real fluids, each particle is subjected *a priori* to the force field of *all* particles (whose effect is quantified by the potential of mean force), whereas in discrete lattice gases with NLIs, each particle interacts nonlocally with at most *one* other particle at a time. So, *stricto sensu*, the usual concept of intermolecular potential does not apply to lattice gases. In the LGA with NLIs, the idea of an interaction range is introduced by governing the interaction distance according to a probability distribution $p(r)(\propto r^{-\mu})$. For sufficiently long times and large number of particles, the implementation of a probability distribution of interaction distances has a resulting effect similar to an effective interaction potential. In order to define a quantity that can be identified as an interaction potential in a discrete lattice gas, one can use the following heuristic argument: The rate of momentum exchange caused by the nonlocal interaction is given by $F(r) = \zeta\kappa_2^2 q(r)$, with $\kappa_2 = f(1-f)$, and where ζ is a numerical factor whose value corresponds to the average amount of momentum transfer. $F(r)$ is then interpreted as a force, from which a "pair potential" can be defined as the discrete analog of the potential in continuum mechanics: $u(r) = -\gamma\mathcal{F}(r)$, where $\mathcal{F}(r)$ is the repartition function corresponding to the distribution $q(r)$, which is directly related to $p(r)$. So $u(r)$ is well defined once $p(r)$ is fixed. For instance, with a power-law distribution $p(r) \propto r^{-\mu}$ such that the interactions are repulsive for $r = 1$ and attractive for $r = 2, \ldots, r_{\max}$, $u(r)$ exhibits a form compatible with the expected typical pair interaction potential of simple fluids.

The static structure factor $\rho S(k) = \langle \sum_{i,j} \delta n_i^*(k,t)\, \delta n_j(k,t)\rangle$ is a constant in the ideal gas, and so it is in the ideal lattice gas: $S_0(k) = (1-f)$, reflecting the absence of spatial density correlations [the factor $(1-f)$ arises because of the exclusion principle]. Now, for LGAs with NLIs, $S(k)$ should be of the form $S(k)/S_0(k) = 1 + fh(k)$, where $h(k)$ is the Fourier transform of the pair correlation function $[g(r)-1]$ and is therefore related to the potential of mean force $\phi(r)$ since $g(r) = \exp[-\beta\phi(r)]$ (here β is an arbitrary constant). Thus, by measuring the density fluctuation correlations in lattice gas simulations, one can extract a function $\phi(r)$ from the measured static structure factor. Figure 10 shows that both the radial distribution function $g(r)$ and the potential function $\phi(r)$ resemble those obtained in real fluid measurements.

At the macroscopic level, LGAs with NLIs can exhibit spinodal decomposition, and in the appropriate density range, one can "quench" the system by increasing the interaction range $r_{\max}$. Then $S(k)$ measured at increasing values of $r_{\max}$ grows dramatically at low k. Using the expression for the compressibility χ of the lattice gas, one has

$$S(k \to 0) = \rho\beta^{-1}\chi \simeq \frac{1-f}{1-\zeta\kappa_3\langle r\rangle_q}; \qquad \kappa_3 = f(1-f)(1-2f) \tag{81}$$

where $\langle r\rangle_q$ is the expectation of r computed with the distribution $q(r)$. Since $\langle r\rangle_q$ increases with $r_{\max}$, we see that $S(k\to 0)$ grows accordingly. The increase of $S(k)$ at low k is characteristic of the amplification of long-range correlations near a phase transition.

The Boltzmann approximation neglects all mode-coupling effects. A consequence of mode coupling is that time correlation functions generally exhibit long-time behavior usually in the form of algebraic decay: $\sim t^{-d/2}$, where d is the space dimension. The existence of these *long-time tails* implies that in dimensions less than or equal to 2, the hydrodynamic equations are valid only for regimes in which mode-coupling effects are negligible, and in dimensions 3 and higher, the form of the hydrodynamic equations remains valid, but the transport coefficients are renormalized.

In mode-coupling theory (see Section VI), one starts with the idea that the long-time behavior can be explained on the basis of hydrodynamic arguments. Consider the case of the velocity of a tagged particle; the mode that describes the decay of its velocity correlations, the shear mode, and the mode that describes particle displacements, the diffusion mode, are coupled. The assumption is that eventually the particle velocity will be equal to the fluid velocity, so that the velocity of the particle is expressed in terms of the particle probability density and of the fluid velocity fluctuations. The former obeys the diffusion equation and the latter the linearized Navier–Stokes equation. So the basic assumption combines the solutions of the two equations, and the result for the normalized velocity autocorrelation function $\psi(t)$ reads

$$\psi(t) \simeq \frac{d-1}{d}\frac{1}{n}[4\pi(\nu + D_s)t]^{-d/2} \tag{82}$$

where n is the number density, and D_s and ν denote the self-diffusion and the kinematic viscosity coefficients, respectively. The same result holds for lattice gases with $n = \rho/v_0$, the number density per elementary unit volume of the lattice (e.g., $v_0 = \sqrt{3}/2$ in the triangular lattice), and with an additional factor $(1-f)$ because of the exclusion principle.

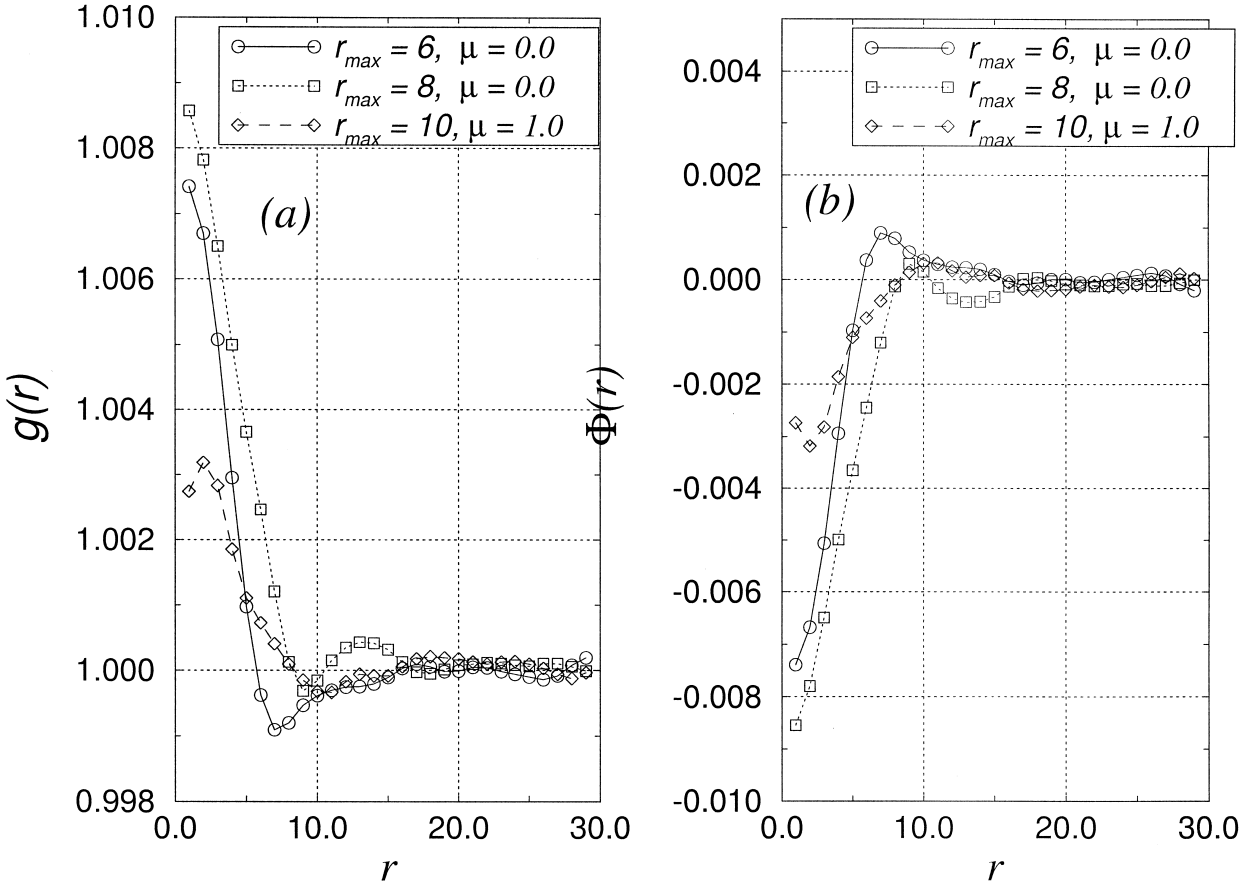


FIGURE 10 (a) Radial distribution function $g(r)$ and (b) potential function $\Phi(r) \equiv \ln(g(r))$ of the lattice gas with nonlocal interactions governed by the probability distribution $p(r) \propto r^{-\mu}$ for $1 \leq r \leq r_{\max}$; $r_{\max} = 6$, $\mu = 0$ (circles), $r_{\max} = 8$, $\mu = 0$ (squares), $r_{\max} = 10$, $\mu = 1$ (diamonds). Lines are guides to the eye.

In order to compute the velocity autocorrelation function of a particle in the lattice gas where all particles are indistinguishable, one must be able to follow the dynamics of a given particle. For this purpose, a very efficient procedure was developed so that high computational accuracy could be achieved for the detection of long-time tails, which requires high precision. Simulations were performed for 2D and 3D lattice gases, and the decay of the computed velocity autocorrelation function $\sim t^{-d/2}$ was found to be in agreement with the mode-coupling prediction; extended mode-coupling theory improved these results to almost perfect agreement with the simulation results for the amplitude factor of the long-time tail]. The evidence of the algebraic decay $\sim t^{-d/2}$ of the velocity autocorrelation function as demonstrated by lattice gas simulations constitutes by far the most accurate verification to date of mode-coupling predictions for hydrodynamic long-time tails: It is one of the convincing achievements of the lattice gas approach to statistical mechanics, showing that LGAs are well suited to serve as a testing ground for concepts in kinetic theory.

ACKNOWLEDGMENT

This work has been supported by the National Science Foundation under Grant CHE-8806767 and by the Fonds National de la Richerche Scientifique (FNRS, Belgium).

SEE ALSO THE FOLLOWING ARTICLES

FLUID DYNAMICS • HYDRODYNAMICS OF SEDIMENTARY BASINS • LIQUIDS, STRUCTURE AND DYNAMICS

BIBLIOGRAPHY

Boon, J.-P., and Yip, S. (1980). "Molecular Hydrodynamics," McGraw-Hill, New York, and Dover reprint edition (1991).

Ernst, M. H. (1990). Statistical mechanics of cellular automata fluids. *In* "Liquids, Freezing and the Glass Transition" (J. P. Hansen, D. Levesque, and J. Zinn-Justin, eds.), North Holland, Amsterdam.

Götze, W. (1990). Aspects of structural glass transitions. *In* "Liquids, Freezing and the Glass Transition" (J.-P. Hansen, D. Levesque, and J. Zinn-Justin, eds.), North-Holland, Amsterdam.

Kim, B., and Mazenko, G. F. (1990). Fluctuating nonlinear hydrodynamics, dense fluids, and the glass transition. *Adv. Chem. Phys.* **78,** 129.

Martin, P. C. (1968). "Measurements and Correlation Functions," Gordon and Breach, New York.

Richter, D., Dianoux, A. J., Petry, W., and Teixeira, J. (eds.) (1989). "Dynamics of Disordered Materials," Springer-Verlag, Berlin, contributions by L. Sjögren and W. Götze, F. Mezei, and W. Knaak.

Rivet, J. P., and Boon, J. P. (2000). "Lattice Gas Hydrodynamics," Cambridge University Press, Cambridge, U.K.

Yip, S. (1979). Renormalized kinetic theory of dense fluids. *Ann. Rev. Phys. Chem.* **30,** 547.



Musical Acoustics

A. H. Benade
Case Western Reserve University

I. Introduction
II. The Plucked and Struck String Instruments
III. The Singing Voice
IV. The Wind Instruments
V. The Bowed String Instruments
VI. The Aptness of Instrumental Sounds in Rooms

GLOSSARY

Harmonic signal Signal whose sinusoidal components have frequencies that are (any) integer multiples nf_0 of some "fundamental" frequency f_0. The repetition rate of such a signal is f_0, whether or not the component f_0 is itself present.

Heterodyne components "Crossbred" components having frequencies $f_v = mf_a \pm nf_b \pm \cdots \pm \cdots (m, n, \ldots$ integers) that are present in the response of a nonlinear system when it is driven by the frequencies f_a, $f_b, \ldots$.

Impedance Ratio F/v of the oscillating excitory force F to the resulting velocity response v at some specified point in a system that is driven at the frequency f. The ratio governs all interactions of the system with whatever is connected to it at its driving point, and it has maxima (or minima) at the system's modal frequencies.

Inharmonic signal Signal whose component frequencies are not integer multiples of a common basis f_0.

Linear system System whose net response to a superposition of stimuli is the sum of the responses to each taken separately, with its response spectrum including only the frequencies present in the stimulus spectrum.

Modal frequencies "Natural frequencies" characteristic of the vibrations of a complex system when it is struck and allowed to ring. The structure of the system determines these frequencies in a unique way (see mode of oscillation).

Mode of oscillation One of a set of distinct ways that a finite linear system moves when it is impulsively disturbed, with each member of the set having its own modal frequency, decay rate, and vibration shape.

Nonlinear system System whose response to a set of stimuli is not a simple superposition of the effects of the stimuli, and whose response spectrum therefore includes heterodyne frequencies in addition to those present in the original stimuli.

Precedence effect Ability of the auditory system to combine several signals arriving in close succession into a single detailed percept, based on all the information that has been collected.

Radiation Transmission of vibrations from a structure into the adjacent (approximately unbounded) region, as from a piano string into its soundboard, or from a trumpet into the concert hall.

Resonance Generic name for the selectively strong

response of a modal system when it is driven by a sinusoidal force whose frequency is close to one of the system's modal frequencies.

Room average The result of averaging a source's signal spectra observed at many points in a room, which gives a useful measure of the sound spectrum of the source itself. Averaging is needed because the transmission of sound between two points in a room is a chaotic function of both position and frequency.

Spectrum A listing of the frequencies and amplitudes of the sinusoidal components making up any signal.

Vibration shape The characteristic distribution throughout a system of the motion of each mode of its vibration, with all parts of the system moving synchronously at the corresponding modal frequency, if this mode alone has been excited.

MUSICAL ACOUSTICS is the scientific study of the arts of performance and composition and the craft of instrument construction as these interact in the context of music. The scientific disciplines most directly involved are oscillation physics and perception psychology, though the muscular and neurological branches of physiology also make important contributions. The emphasis of the present article is on the dynamical properties of musical instruments as these are influenced by the environment in which they are played and by the needs of the listener's auditory mechanisms.

I. INTRODUCTION

This article on musical acoustics opens with an outline of those salient properties of the human auditory system which permit it to function as a processor of musical sounds in the concert hall. The apparently chaotic properties of the sound-transmission process in halls will also be sketched, along with an indication of how the ear can collect the data it needs undistracted by the confusion. Even though the article concerns itself primarily with the physical behavior of musical instruments and musical sounds, it is important for the reader to begin with a good idea of the perceptual requirements that instruments must satisfy if they are to be musically successful. The physical properties of musical sounds and of the transmission path from source to listeners are complicated. Music is possible because its sounds are of a type whose significant features may readily be deduced from the received signals. For signals of the proper type, the auditory processor proves to be an extremely efficacious signal processor that performs so effortlessly that few people are even remotely aware of the complexity of its task.

The following assertions will serve to provide a framework for the perception parts of this initial outline.

1. The mechanical motions of a primary sound source give rise to an acoustical signal that is ultimately processed by the listener's neurophysiological system.
2. Sounds from a source normally come to a listener's ears via a set of multiply reflected transmission paths in a room.
3. The listener's auditory system selects one or more subsets of the signal information coming to his ears, and he uses these in mutually supportive ways to recognize features of interest. The subsets that come via different transmission paths are not always equivalent.
4. The system normally has several modes for processing and recognizing any given signal feature. It is then able to resolve perceptual ambiguities, choosing those modes that give consistent results while "setting aside" the others.
5. Much of the recognition function of the auditory system is based on sorting the signal properties into categories associated with the characteristic behavior of generic types of musical sound sources. Because of this, even partial information is often sufficient to distinguish between instruments and for following individual instrumental voices in an ensemble.
6. The musically important recognition processes are all perceptually robust. That is, they are based on properties of the musical signal that not only survive the transmission path but also are insensitive to the presence of other musical signals or noise.

A. Structural Sketch of the Auditory Detection System

We will pass over the outer ear (a sound collector) and the middle ear (a coupling device and first overload protector) to focus our attention on the major functional part of the inner ear, the cochlea. The basilar membrane within the cochlea provides the preliminary frequency sorting device in the auditory signal path, as well as the transduction equipment that encodes mechanical vibrations into nerve impulses for further processing. The sorting function of the cochlea is very simple: if a sound made up of several sinusoidal components of widely separated frequencies enters the ear, each one produces its own localized disturbance at some point along the basilar membrane—low frequencies at one end, high at the other. Thus, the firing of receptors at a particular position along the membrane indicates that a certain frequency is present in the original sound.

This sorting by place along the basilar membrane is not fine-grained. The perceived correlate (pitch) of the

vibrational frequency of a sinusoid is associated with the place of maximum vibration of the basilar membrane, while the maximum implied by the response gradient in the "skirts" of the response region tends to define it with greater precision.

When two stimuli whose frequencies f_a and f_b differ by less than about 20% enter the ear, their disturbances overlap on the basilar membrane, and so stimulate the same set of receptors. The two stimuli produce cyclic alternations in the vigor of the local vibration as they run in and out of step with one another, with an alternation (mechanical beating) rate equal to the difference $f_a - f_b$ between the stimulus frequencies. On the other hand, when the acoustic stimulus frequencies are separated by much more than about 25%, two essentially unrelated and mechanically noninteracting sets of receptors transmit data about the two stimuli.

A very large fraction of auditory processing theory is shaped by the existence of critical bands whose ultimate origin lies in the distributed response of the basilar membrane (although modified by a certain amount of near-neighbor interaction of the receptors themselves). Over most of the frequency range of musical interest, each critical band extends over a range of roughly 26% ($\frac{1}{3}$ octave).

Because of its importance in musical listening, we should examine a few examples of the role of the critical band phenomenon, beginning with the simplest. When the ear is presented with two closely spaced sinusoids, such that $f_a - f_b \leq 20$ Hz (well within a critical band), the listener directly perceives the mechanical pulsations of the basilar membrane as a pulsation in loudness of a sinusoidal signal, whose pitch belongs with a frequency of about $(f_a + f_b)/2$. This is exactly in accordance with the expectations of a physicist. However, when $f_a - f_b \geq 20$ Hz, the signal is not perceived as having a rapidly pulsating loudness, but as a rather rough sound instead. The roughness of this sound decreases rapidly however as the frequency difference is increased toward the extent of the critical band (e.g., to about 114 Hz for sinusoids lying near the note A_4, whose repetition rate is 440 Hz).

The determination of loudness is another auditory process that is dominated by the critical band phenomenon. The perceived loudness of a sound having all of its acoustical energy E concentrated in a single critical band varies very nearly as $E^{0.3}$ for stimuli within the musically important range of signal levels. This functional relationship holds unchanged whether all the signal energy is carried by a single sinusoid or by a group of closely spaced ones, or even when the critical band is filled by random noise. When, however, two or more widely separated critical bands are provided with stimuli $E_1, E_2, \ldots$, the total loudness is perceived as the simple sum of the loudnesses contributed by each in its own right. For components of intermediate spacing, allowance must be made for the fact that the band edges are not sharply marked, so that they "shade off" from one to the next. In any event, the loudness will be greater when a given amount of power is apportioned among several critical bands than when it is concentrated within a single critical band.

We now examine how the vigor of acoustical signals is neurologically encoded. The basilar membrane is richly supplied with receptors all along its length (some 1500 per critical band). If the local vibration at the position of one of these receptors is sufficiently strong (to first approximation), it fires once per cycle of the vibration, sending an electrical pulse to the higher centers of neurological processing. These receptors have a wide variety of threshold sensitivities, so that only a few can be fired by a weak local vibration, while many of them produce pulses when the vibration is strong. In the language of signal processing, we may say that the signal frequency is coded in part by its location along the cochlea and in part by the repetition rate of neural firing, whereas the signal strength is chiefly represented by the number of receptors that fire in each burst (or volley, as it is customarily called).

Once a receptor has fired, it becomes insensitive for a refractory period that lasts about 1 ms, after which its sensitivity rapidly returns to normal. For sinusoidal stimuli having a frequency of 2000 to 3000 Hz, a given receptor will fire (on the average) only on every second or third cycle of the stimulus. For a given strength of mechanical disturbance in the cochlea then, the average number of receptors actually firing per cycle is less for high-frequency stimuli than for those occurring below about 1000 Hz.

This last statement may lead us to speculate that the ultimate perception of musical sounds may well be different for tones that have a majority of their sinusoidal components below 1000 Hz and those having a preponderance of high-frequency components. There are a number of familiar properties of musical sounds that support such speculations. For the moment, we need only to suggest the simplest of them all: the vast bulk of musical composition worldwide is written to encompass a pitch range from a few notes below C_4 (whose repetition rate is 261 Hz) to a few notes above C_6 (repetition rate 1044 Hz). Furthermore, we recognize the existence of many instruments and many musical parts that are pitched in regions of the musical scale that are very much lower than this "musical heartland," while there are very few to be found in the higher-pitched musical regions. We shall return to these matters in Section VI of this article.

Communications engineers who make use of pulse coding for their signals will recognize that the threshold behavior of the auditory receptors and the subsequent and-gate and or-gate behavior of higher-level synapses joins with the "hit one, miss a few, and hit another"

responses of the primary receptors to generate pulse trains of frequency f_0 at many points in the nervous system network, and regardless of the details of its structure if some of the receptors are stimulated by incoming sinusoids whose frequencies are integer multiples of $\alpha f_0, \beta f_0, \gamma f_0$ of the same fundamental repetition rate. For example, an ear provided with signals having the frequencies 600, 800, and 1000 Hz will (among other things) give rise to a widespread neurological pulse-rate $f_0 = 200$ Hz. On the other hand, sounds containing collections of inharmonic components ($\alpha, \beta, \gamma, \ldots$, where α, β, and γ are not integers) give rise to essentially chaotic pulse trains in the nervous system, although subsets of these tend to coalesce to give nearly periodic but fluctuating volleys if members of the subsets are approximately harmonic. We find that the existence and nature of many fundamental properties of musical structure are suggested (if not implied) by this global property that results from the mere existence of synaptic action. In particular, we may use it as a clear hint of why the ear responds to a sound made up of a collection of harmonic partials as a single clear perceptual entity—the musical tone—while an unrelated collection is perceived as a jumble of separately heard sounds.

The fact that each one of several simultaneously presented harmonic complexes is clearly heard as a tone in its own right is also strongly hinted at by the same properties of the pulse coding action. Obviously, integer relations between the repetition rates of these tones will themselves be expected to produce perceptually significant phenomena of the sort that underlie formal music theory.

B. Sound Transmission in a Room

It has already been remarked that the signal path between the sound source and receiver in a room is highly variable, and even chaotic. Because experience shows that performers and listeners alike find it easier to carry on their musical activities in a room than in a reflection-free environment, we must first learn something of the physical nature of the acoustical transmission path and then outline a few of the neuropsychological methods used by the listeners to musically exploit information gained via this path.

Consider a prototype experiment in which an oscillator sinusoidally drives a loudspeaker at one point in a room. Let the oscillator frequency be slowly raised, while a microphone placed at another point in the room has its output signal traced out on the moving paper of a strip-chart recorder. Figure 1 shows the resulting record of the transmission of sound from the loudspeaker to each of three different microphone positions in a laboratory room. The first feature of these tracings to catch the eye is their extreme irregularity; we further note that the three traces are entirely different. If, as a matter of fact, we were to produce 30 or 40 traces of this type using various randomly chosen positions for the sound source and the detector, the traces would all be different. However, the mathematical average of the various microphone signal curves does have a well-defined meaning—it serves to inform the physicist of the strength of the loudspeaker signal itself. The sloping dashed line shown in the figure indicates such a room average, and shows that the loudspeaker in the present experiment produced a steadily increasing excitation of the room as its frequency was progressively raised.

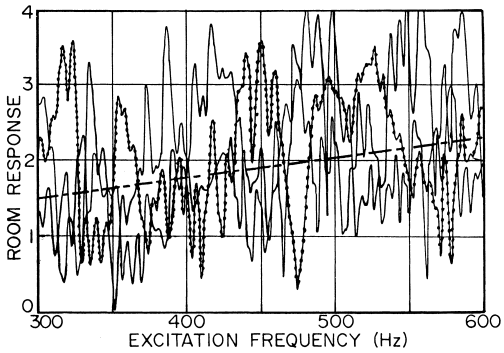


FIGURE 1 Traces showing the extreme irregularity of sound transmission in a room.

Room-average spectra obtained from musical instruments are well-defined, and they are enormously informative about instrumental behavior, but only as long as the player is assigned a familiar performance task adequately specified in musical terms. Note also that, contrary to the belief of some recording engineers, combining all the microphone signals at the input of a single analyzer accomplishes nothing more than a reshuffling of the musical cards. The result has no more significance than does the analysis of a single microphone!

C. Auditory Processes in a Large Room

We have learned that the transmission of sounds in a room is (from the point of view of the physicist) complex and randomly varies from point to point. Let us therefore inquire about one of several methods of signal processing used by the auditory system that permits it to make "measurements" of the arriving sound at a rate that far exceeds the abilities of a scientist using his best equipment.

First consider the signals received by the listener's ear during the commencement of a musical sound in a room. He is provided with a series of early reflections. The first to arrive is the direct sound from the instrument, and then in quick succession come the first reflections from the walls, the floor, and the ceiling of the room. If we were using our eyes in a mirrored room instead of our ears, we could say that these initial reflections provide us with

front, left-hand, and right-hand views, plus information from above and below. In both the acoustical and the optical versions we find reflections of reflections, these becoming ever weaker and less well defined as the various complexities of the reflection and transmission process take their toll. Then we come to an important fact: A few early reflections present the listener/observer with nearly complete physical information about all aspects of the signal source.

In a mirrored room, we would have to shift our conscious attention from one available view to another, and then to intellectually combine the gathered information into an overall picture of the observed subject. The musical listener, on the other hand, can perform his compilation of data in what is perceived to be zero time! The successful collection, storage, comparison, and interpretation of these early-arrival data is possible if the successive early reflections bring in their messages within a time interval of about 35 ms of one another. This implies that the reflecting surfaces must be sufficiently close to either the source or the listener that most of the sounds arriving from them have traveled no more than about 10 m farther than their predecessors. It is impossible to overemphasize the importance of this fact! We have here the essential clue as to how the auditory system carries on a major part of its work in the concert hall. There is one more feature of the signal processing behavior of the musical ear to which we must give our serious attention: reflections that have traveled distances of more than about 40 m farther than the direct sounds are actively disruptive of the musical recognition process (see Fig. 2). The general nature of this type of auditory processing has been recognized for over 150 years and empirically exploited for a century. Its basic manifestations were scientifically formalized in the 1950s under the name precedence effect.

From the point of view of a physicist, the buildup of sound at some place in a room is produced by the superposition of the direct sound and a set of successive reflections that arrive with random phase, an ever-increasing arrival rate, and progressively weakening amplitudes. Clearly, the buildup of sound (and its analogous reverberant decay) is itself a violently fluctuating random process. Just as averaging many samples of steady-state sound in a room gives a useful room average p_{avg}, so also does the averaging of many buildups or decays lead to a measure (the only meaningful one) of the reverberation time T_{r} of the room.

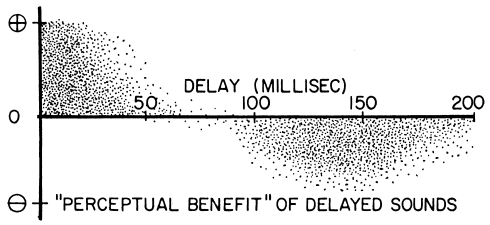


FIGURE 2 Reflections arriving within about 35 ms of the original sound enhance the listener's perception of musical sounds in many ways. Later arrivals degrade these perceptions.

Taking the two averages p_{avg}, T_{r} together, we may then say that the average onset behavior $p_{\text{onset}}(t)$ of the room sound follows the rule

$$p_{\text{onset}}(t) = p_{\text{avg}}\left(1 - e^{-6.91t/T_{\text{r}}}\right), \tag{1}$$

while the decay is given by

$$p_{\text{decay}}(t) = p_{\text{avg}}\, e^{-6.91t/T_{\text{r}}}. \tag{2}$$

We must not forget, however, that it is only the earliest 50 or 60 ms of the onset that contribute to fine-grained musical detection, while the remainder provides little more than the "aroma" of earlier sounds.

When summarizing the foregoing discussion of the functioning of the human auditory system in the concert hall, we begin by emphasizing that while the ear can collect data over several tens of milliseconds, these early reflections are fused into a single percept. The time of occurrence of the percept turns out to be the instant at which the earliest contribution arrives, which is normally the arrival time of the direct sound from the instrument. Also, it is perceived as coming from the point in the room from which the instrument itself transmits the first-arriving signal, and its loudness is perceived as being accumulated from the entire sequence of early arrivals. The individual arrivals are used together to provide a mutually confirmatory basis upon which we can assess such musical features as tone color, stability of production, and the type of articulation chosen by the performer for the note under consideration. It is quite correct to say that the ear is able to deduce the room-average sound by a suitable processing of the room-caused fluctuations in its onset (or decay).

II. THE PLUCKED AND STRUCK STRING INSTRUMENTS

A. The Guitar

The guitar will be used as a prototype instrument to trace the major relationships that adapt the physical structure of a musical instrument to the auditory requirements of the musical ear. As indicated in Fig. 3, the generic guitar may be described as consisting of a set of six strings supported on a hollow, thin-walled body and a fairly rigid neck. When a string is plucked, its vibrating length extends from an anchorage on the bridge (attached to the top plate of the body) to one of the frets on the neck (as selected by the player). The instrument is of course normally played in

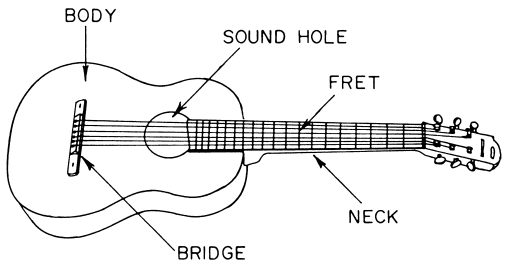


FIGURE 3 Diagram of a guitar showing the names of its major parts.

a room or concert hall. The physicist recognizes from this listing that he is dealing with a coupled sequence of one-, two-, and three-dimensional vibratory systems. Taken in aggregate, such a sequence has an enormous number of modes to be excited at one point (on a string) by a player and observed (in the room) by the listener. The characteristic impedance of the string subsystem is much smaller than that of the wooden structure to which it is attached, while the characteristic impedance of the body is enormously greater than that of the air-filled room. Because of these differences, the three subsystems interact weakly, and their modal properties may usefully be analyzed one by one.

The string itself is a one-dimensional system. Plucking it excites a set of modal oscillations whose frequencies are arranged in an almost precise whole-number relation as

$$f_n = (n/2L)(T/\mu)^{1/2}. \tag{3}$$

Here L is the string length, T its tension, and μ its mass per unit length. However, the exactness of this harmonic relationship is slightly disrupted by a small correction arising from the barlike stiffness of the string and by irregular perturbations associated with its coupling to the guitar body.

Because the modal frequencies of a guitar string enjoy a harmonic relationship, the instrument transmits to the listener sounds having frequency patterns that his neurophysiological processor responds to with particular vividness, and which govern the formal structure of music.

The vibrations of a plucked string drive the guitar body mainly by way of an oscillatory force exerted on the bridge, although there is a musically significant excitation applied by the other end of the string and transmitted to the body via the neck. Once corrections have been made for modal behavior associated with the guitar's barlike neck and the air contained within the instrument's cavity, the mean spacing of the body modes is essentially constant ($\simeq$100 Hz) over the entire frequency range. Mathematical analysis shows that this behavior is to be expected for a platelike object, even when segments of various thicknesses are joined into a boxlike structure.

When the guitar body is excited by the abruptly begun vibrations of a plucked string, its motion can be classified into two separate but musically significant categories. First, the body vibrates as a driven system at the frequencies of the excitory string modes. The vigor of each component motion depends on the point of application of the driving force, and its decay in amplitude is governed by the decay rate of the corresponding string vibration. The sound components produced in the room by these vibrations exert a major influence on the pitch, loudness, and tone color of the guitar's perceived signal. The second aspect of the guitar sound is associated with the body, in which a transient vibration is set up whenever a string is plucked. The frequency components and the decay rates of this sound are those characteristic of the body modes themselves. While there is little energy associated with this part of the sound arriving at the ear, it contributes significantly to the hollow, woody tone color and mild initial thump that is characteristic of all guitar notes.

The ability of the guitar body to convert its string vibration excitation into audible sound in the room depends significantly on the relationship between the wave speed of sound in the body to that in air, on the number and kind of discontinuities existing in the body structure, and on the frequencies and strengths of the modal resonances of the body. However, for an elaborately intertwined set of reasons that will be elucidated little by little throughout the course of this article, the excitory force applied by the string to the guitar bridge is converted into sound in the room with an average efficacy that is almost independent of frequency. To be sure, there are peaks and dips in the radiation processes associated with the guitar body parameters mentioned previously, but (for reasons that will become clearer over the remaining course of this article) these may vary widely from instrument to instrument without destroying the recognizability of its sound as being guitarlike. Only one or two of these peaks of radiative efficacy play a major role in subjective judgments of guitar quality, and these lie in the region of the lowest frequency components of the guitar's lowest notes.

The physics of plucked string motion is such that the amplitudes of the various modes excited by a given type of plucking depend very strongly on the position of the plectrum along the string and on its breadth and hardness. Temporarily setting aside the modifications arising from the nature of the plectrum and the stiffness of the string, the force F_n exerted on the bridge by the nth vibrational mode of a string is

$$F_n = (2F_0/\pi n)\sin(n\pi x_e/L). \tag{4}$$

Here L is the vibrating length of the string, and x_e is the distance away from the bridge that an excitory force F_0 is applied by means of a very narrow plectrum. The presence of the mode number n in the denominator shows that

the high-frequency modes exert progressively less driving force on the bridge (and so produce less sound in the room) than do the lower modes.

When a guitar is played, the plucking point position x_e tends to be roughly constant, while the player chooses any one of a wide variety of string lengths L by his use of the frets. Equation (3) may be used to reexpress F_n as the function $F_a(f)$ that gives the driving force exerted by any string oscillation having the frequency f, whether it be associated with a large-n mode of a long string, or a small-n mode of a short one:

$$F_a(f) \equiv K \cdot E_a(f) = K\left[\frac{\sin(\pi f/f_a)}{\pi f/f_a}\right]. \qquad (5)$$

Here K is $(2F_0x_e/L)$ a proportionality constant, and f_a is the first-mode frequency of the string when it is shortened to make its length equal to the bridge-to-plectrum distance x_e.

We may usefully characterize the function $E_a(f)$ as the spectrum envelope of the bridge driving force produced by a simple plucked string. When a particular string length L_p is chosen by the player, the resulting sound is made up of sinusoids having frequencies that are integer multiples of the first-mode frequency f_p. The resulting force spectrum amplitudes at the bridge are then proportional to $E_a(f)$ evaluated at the frequencies of nf_p.

The left-hand curve in Fig. 4 shows the general shape of the $E_a(f)$ function. There are "spectral notches" in E due to the zeros of the sine function (i.e., frequencies of zero sound production), while the whole curve has its behavior dominated by a solid trend line, which shows that E is essentially constant for frequencies below the breakpoint value $\alpha = f_a/\pi$ and falls with an asymptotic rate $1/f$ for $f \gg \alpha$.

A guitar is not normally excited by a narrow plectrum. The width of the actual plectrum (or fingernail, or finger tip) joins with the inherent stiffness of the string to produce a rounding-off of the string profile near the plucking point, instead of the abruptly angled profile that was assumed for the calculation of Eqs. (4) and (5). This rounding-off of the string profile produces a systematic modification $E_b(f)$ of the guitar's sound spectrum envelope as

$$F(f) = K E_a(f) \cdot E_b(f), \qquad (6)$$

where

$$E_b(f) = \left[\frac{\sin(\pi f/f_b)}{(\pi f/f_b)[1 + (f/f_b)^2]}\right]. \qquad (7)$$

Here f_b is the first-mode frequency of a string whose length w_b is equal to the width of the curved region near the plucking point. The general nature $E_b(f)$ is shown by the middle curve in Fig. 4. Again, the notches associated with the sine function are shown dotted, leaving a solid line to show the main trend of this contribution to the spectrum envelope. Note that E_b is essentially flat below a breakpoint frequency $\beta = (f_b/\pi)$ determined by w_b, while the amplitude falls at an asymptotic rate of $1/f^3$ for frequencies well above β.

If the second breakpoint frequency β is much higher than the first, the overall spectrum envelope starts out independent of frequency at low frequencies, rolls over and begins to fall as $1/f$ for $f > \alpha$, and then falls at the rapid rate of $(1/f)(1/f^3) = (1/f)^4$ above $f = \beta$, where the joint effects of E_a and E_b are active. If α and β are not so very different, there is a transition region of intermediate slope between the level behavior at low frequencies and the $1/f^4$ high-frequency fall-off. Figure 5 shows the results of this behavior, as found in the practical world of guitar playing. Here the bridge force is shown as a function of frequency for a string that is plucked at a fairly normal point $\frac{1}{8}$ of the way along it by a plectrum, whose width gives a local string curvature extending over $\frac{1}{16}$ of the

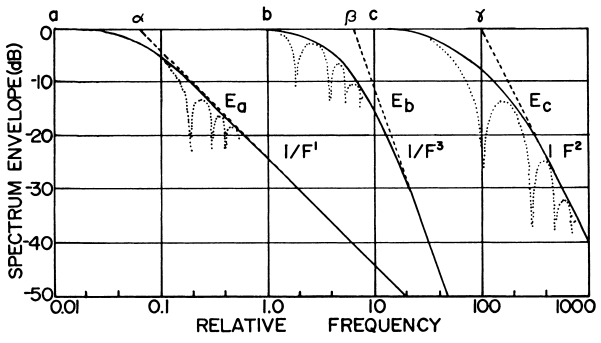


FIGURE 4 $E_a(f)$, $E_b(f)$, and $E_c(f)$ combine to produce the spectrum envelope of guitar tones. E_a is associated with the point of plucking, E_b with plectrum width, and E_c with string stiffness effects.

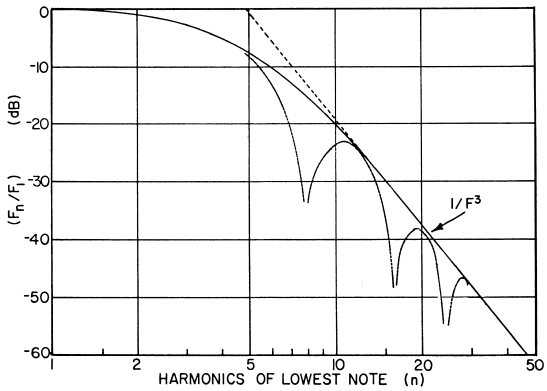


FIGURE 5 Overall spectrum envelope of the guitar bridge driving force. The trend is from constancy at low frequencies to a high-frequency rolloff proportional to $1/f^3$.

"open" (maximum) string length. For reasons that will become clearer as we go along, a line is drawn on this graph to suggest that the high-frequency behavior of the spectrum is well approximated by a $1/f^3$ falloff rate, with a breakpoint located at about five times the open string first-mode frequency. The spectral notches are shown explicitly only for the $E_a(f)$ aspect of the behavior.

As remarked earlier, when a player is performing music on a guitar, he tends to pluck the strings at a roughly constant distance from the bridge. The bridge driving-force spectrum for all the notes played on any one string will then share a single spectral envelope of the form shown in Fig. 5, including the notches. While the spectrum envelopes for the adjacent strings (tuned to different pitches) are alike in form, the notches and breakpoints are displaced bodily to higher or lower frequencies by amounts depending on the tunings of these other strings. Taken as a group, however, the force exerted on the bridge by all the strings has an overall envelope that is frequency independent at low frequencies and varies roughly as $1/f^3$ at high frequencies. The transition region between those two behaviors is blurred somewhat, due to the differences between breakpoint frequencies belonging to the individual strings.

While the gross envelope for the bridge driving-force spectrum is in fact made up of six distinct parts (one for each string), the mechanism for conversion of this excitation into the room-average sound is very much the same for all strings. It is all mediated by the same set of body resonances, and these can produce fluctuations in the radiated sound above and below the essentially frequency-independent trend line of the overall radiation process for sounds emitted by a platelike object. Thus the observed room-average spectrum is found to have fluctuations above and below an overall envelope whose shape is very similar to that of the curve in Fig. 5. It is easy to estimate the expected number of fluctuations over any frequency span of interest. This is equal to the number of body modes found in this span, augmented by the corresponding number of notches in the drive-force spectrum.

It is appropriate here to inject an additional piece of information about the human auditory processor. A music listener is (in a laboratory situation) readily able to detect the strengthening or weakening of one or more sinusoidal components of a harmonic collection. However, in the processing of music or speech he does not "pay attention" very much to the presence of holes or notches in the spectra of the sounds that he processes as a means for recognizing them or assessing their tone color. More precisely, failure to provide notches will be noticed and criticized in an attempted sound synthesis, but (except for certain special cases) their mere presence and their mean spacing are of more significance than their exact positions.

B. The Harpsichord and Piano

The harpsichord and piano are acoustically similar to the guitar in that they have a set of vibrating strings that only gradually communicate their energy to a two-dimensional platelike structure (the soundboard), which in turn passes on the vibration in the form of audible sound to the room. Once again, for musico-neurological reasons, it is important that the primary string mode oscillations take place at frequencies that are in whole-number relation to one another. The major difference that distinguishes these instruments from the guitar is the fact that they lack frets, so that each string is used at only a single vibrating length; also, the place and manner of plucking or striking is chosen by the instrument's maker rather than by its player.

The strings of a harpsichord are excited by a set of plectra (jacks) operated from a keyboard, so that in many respects the dynamical behavior of a harpsichord is identical with that of the guitar. As a result, the main features of the spectral envelope of harpsichord sounds in a room are the same as those of a guitar. The curve shown in Fig. 5 and the accompanying discussion apply equally well to the harpsichord, the chief difference being in a slightly different distribution of spectral notches associated with the excitation mechanism, and a greatly decreased mean spacing ($\simeq$20 Hz) of the radiation irregularities that are associated with the plate modes of the soundboard. The (as yet undiscussed) air resonances of the cavity under the harpsichord's soundboard play a very much smaller role in determining the overall tone than is the case for the guitar.

The mechanical structure of the piano is quite analogous to that of the harpsichord, but the use of hammers rather than plectra to excite its strings causes several modifications to the overall envelope function. While it is commonly believed that the striking point should have an effect on the envelope similar to that given by $E_a(f)$ for the plucked string, in fact the corresponding function is much less dependent on frequency, with only small dips appearing at the frequencies of the "notches" of $E_a(f)$. However, the width and softness of the hammer join with the string's stiffness to give rise to an envelope function $E_a(f)$ exactly as given for the plucked strings. There is, however, one more dynamical influence on the spectral envelope of a struck string: When the hammer strikes the string, it bounces off again after a time that is jointly determined by the hammer mass and elasticity, the string tension, the position of the striking point along the string, and the length of the string. The details of this dependency of the hamer contact time on these parameters are complicated, and it will suffice for us to present only its main spectral

consequences. It gives rise to an envelope function $E_c(f)$ that is satisfactorily represented by a simplified formula as

$$E_c(f) = \left[\frac{\cos(\pi f/4f_c)}{1+(f/f_c)^2}\right]. \tag{8}$$

Here f_c is very nearly equal to the reciprocal of the hammer contact time. The nature of this function is shown in the right-hand part of Fig. 4. The envelope function has a familiar form, being essentially constant at low frequencies, having a number of deep notches, and ultimately falling away as $1/f^2$ at high frequencies with a breakpoint γ for the main trend at $f = f_c$. The analog to Eq. (6) for the piano is then

$$F(f) = K \cdot E_b(f) \cdot E_c(f). \tag{9}$$

At very high frequencies ($f \gg \alpha, \beta$, and γ), the drive-force spectrum falls away as $(1/f^3)(1/f^2) = (1/f^5)$. As before, the behavior at somewhat lower frequencies can have an apparent fall rate represented by some intermediate exponent that depends on the maker's choice of β and γ.

In the piano, the distance of the hammer's striking point from the string end varies smoothly (sometimes in two or more segments) from about 10 or 12% of the string length at the bass end of the scale to about 8% at the treble end. Similarly, the mass, width, and softness of the hammers fall progressively in going up the scale from the bass end. Taken together, these four varying parameters of piano design provide the maker with his chief means for achieving what he calls a "good" tone for the instrument simultaneously with uniformity of loudness and of keyboard "feel."

The hammer mass has a direct influence on the feel of the keys. It also plays a major role in determining the loudness of the note via its effect on the kinetic energy that it converts into vibrational energy of the string. The hammer mass (as well as its softness to some extent) joins with the striking point and string parameters to control the contact time during a hammer blow. At C_2 near the bottom of the scale, the design is such that the hammer mass is about $\frac{1}{30}$ of the total string mass, and the hammer's contact time is around 4 ms ($\gamma \simeq 250$ Hz); at the midscale C_4, the string and hammer masses are about equal, and the contact time is about 1.5-ms; at C_7, near the top of scale, it has fallen to 1 ms ($\gamma \simeq 1000$ Hz). In a related manner, the string stiffness joins with hammer softness to determine the string-curvature envelope $E_b(f)$ and its corresponding breakpoint frequency β.

Figure 6 shows the remarkable uniformity in the trend line for the measured room-average spectra (defined in Section I) of notes taken from the musically dominant midrange portion of a grand piano's scale. These notes, running scale-wise from G_3 up to G_5 (having repetition

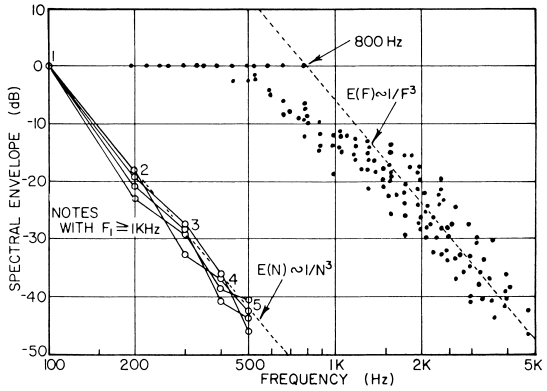


FIGURE 6 Measured room-average spectrum envelope of piano tones. Above about 800 Hz the components weaken as $1/f^3$.

rates 192 to 768 Hz), will be recognized as lying in the region in which the auditory processor is particularly quick, precise, and confident. As a result, any regularities shown by the spectra of these notes have strong implications about the manner in which the ear deals with such notes. In the figure, the dots located along the zero-decibel line represent the normalized amplitudes of the first-mode frequency components of all 14 played notes. The remaining dots then give the relative amplitudes of the remaining higher harmonic components (expressed in decibels relative to the "fundamental" components). About half of the notes shown here were played and measured several times, over a period of five years, using a variety of analysis techniques. The variability due to all causes (irregularity of striking the keys, statistical fluctuations of the room measurement, wear of the piano, and differences due to altered analysis technique) may be shown to be about ± 2 dB for the position of any one dot on the curve. For this reason it is possible to attribute the observed scattering of the points about their basic trend line almost wholly to the excitatory spectrum notches and to the radiation effects of soundboard resonances in the body of the piano. The magnitude of these fluctuations is consistent with estimates based on the resonance properties of a sound board.

Figure 6 illustrates a spectral property that is shared by nearly all of the familiar midrange musical instruments. Here, as in the guitar and harpsichord, we find fluctuations about an essentially constant low-frequency average trend, plus a rolloff with a $(1/f^3)$ dependency at high frequencies. Dividing the two spectral regions, there is also a well-defined break point that lies close to 780 Hz for the piano.

To recapitulate, the proportioning of the piano's string length and strike point and its hammer's breadth, mass, and softness cause the critical frequency parameters f_b and f_c to vary widely for strings over the midrange playing scale. Nevertheless, the maker has arranged to distribute them in

such a way as to preserve the absolute spectral envelope over a wide range of playing notes. Since it is clearly not an accident that the piano and harpsichord have developed with proportions of the type implied previously, we are led to inquire as to what are the perceptual constraints imposed on the design by the needs of the listner's musical ear. The answers to this inquiry (implied in large measure by the auditory properties outlined in Section I) will be made more explicit in the remaining course of this article.

C. Radiation Behavior of Platelike Vibrators

A number of significant musical properties of the guitar, harpsichord, and piano have been elucidated by an examination of the ways in which their platelike body or soundboard structures communicate the vibrations imposed on them by the string to the surrounding air in the room (violins of course also communicate this way). It has already been asserted that the trend of radiating ability of such structures driven by oscillating forces is essentially independent of frequency except at low frequencies. Because a number of musical complexities are associated with this apparently simple trend, it is worthwhile to devote some space to a brief outline of the radiation physics that is involved.

To begin with, consider a thin plate of limitless extent, driven at some point by a sinusoidal driving force of fixed magnitude F_0 and variable frequency f. Analysis shows that the vibrational velocity produced at the driving point of such a plate is proportional to the magnitude of the driving force, but independent of its frequency. For definiteness, let the plate be of spruce about 3-mm thick (as is the case for the guitar and violin top plate, or the soundboard of a harpsichord). Also, we temporarily limit the driving frequency to values that lie below about 3000 Hz.

Despite the fact that the entire surface of the plate is set into vibration by the excitory force, only a small patch near the driving point is actually able to emit sound into the air! The radius r_{rad} of this radiatively effective patch is about 16 cm at 100 Hz, and it varies inversely as the square root of the frequency. Thus, the area of the active patch varies as $1/f$. Since the radiation ability of a small vibrating piston is proportional to its area, velocity amplitude, and vibrational frequency, the sound emitted by the board not only comes from a tightly localized spot at the point of excitation, but also the amount radiated is entirely independent of frequency.

When the size of the plate is restricted by any kind of boundary (free, hinged, or clamped), additional radiation becomes possible from a striplike region extending a distance r_{rad} (defined previously) inward from these boundaries. Any sort of rigid blocking applied at some point, or hole cut in the plate, also gives rise to a radiatively active region of width r_{rad} around the discountinuity, and the system retains its essentially frequency-independent radiating behavior. The fact that the system is now of finite extent means that it has a large number of vibrational modes (whose mean spacing is set mainly by the thickness and total plate area of the structure). The system's net radiated power then fluctuates symmetrically above and below the large-plate trend line in a manner controlled by the size, width, and damping of the modal response peaks and dips. Curiously enough, the general level of radiation is hardly influenced by the plate damping produced by its own internal friction.

Above a certain coincidence frequency f_{coinc} (the previously mentioned 3000 Hz for a spruce plate 3-mm thick), the entire vibrating plate abruptly becomes able to radiate into air. For a limitless plate, the radiating power becomes enormous just above f_{coinc}, and it then drops off to a new frequency-independent value that is considerably greater than that found below f_{coinc}. However, for a finite-sized system broken up into many parts (as in the musical structures), there is no readily detectable alteration in the overall radiating ability as the drive frequency traverses f_{coinc}, although many details of the directional distribution of sound are drastically changed.

The coincidence frequency is inversely proportional to the plate thickness, and the radius r_{rad} of the radiatively active regions is proportional to the square root of the thickness; this means that for a piano (whose plate thickness is about triple that of a harpsichord), f_{coinc} falls to about 1000 Hz and r_{rad} is about 27 cm at 100 Hz.

The practical implications of the radiation properties briefly discussed here are numerous. To begin with, it should be clear that the boxlike (and therefore irregular) structure of the guitar and violin joins with the sound holes and miscellaneous internal bracing to greatly increase the sound output from what would otherwise be very soft-voiced instruments (as are the lutes and viols of simpler construction that were developed earlier).

By the beginning of the seventeenth century, the harpsichord soundboard had already accquired numerous heavy struts along with structural discontinuities provided by the bridges needed to serve two complete sets of strings (the so-called 4- and 8-ft arrays) and a hitch-pin rail between the bridges to bear the tension of the 4-ft strings. These strong and heavy discontinuities play an important role in providing the free, clear sound that is characteristic of a really fine harpsichord. The instruments by the French builder, Pascal-Joseph Taskin (1723–1793), which are noted for the fullness of their tone, are provided with an unusually rigid set of bracings. It is important to notice that despite the completely counterintuitive nature of the lumpy and discontinuous structures that favor sound production, the best makers nevertheless discovered

and adhered to designs whose underlying acoustical virtues have only been elucidated in the late twentieth century.

A quick look at the modern piano shows a similar adaptation of the vibrating structure to its radiating task, although we might today ask (by analogy) whether or not a few properly placed extra braces might improve the sound somewhat.

Some of the difficulty often faced by sound engineers when making recordings of the piano is readily understood in terms of the ever-shifting patchwork of sources that are active. The tendency of many engineers to place their microphones close to the piano means that rapidly changing, and often perceptually conflicting, signals are registered in the two recording channels. The ear is so accustomed to assembling the sounds from all over the soundboard, via the mediation of a room, that anything else can confuse it. As a matter of fact, the ear is so insistent on receiving piano sounds from a random patchwork of shifting sources that successful electronic syntheses of piano music can be done using the simplest of waveform sources as long as signals are randomly distributed to an array of six to ten small loudspeakers placed on a flat board!

D. Piano Onset and Decay Phenomena

When a hammer strikes a piano string it is subjected to an impulsive blow that contains many frequency components. This blow, which is transmitted to the bridge in the form of a continuously distributed drive-force spectrum whose shape is exactly the same as the spectrum envelope of the discrete string vibration drive forces, is heard in the room as a distinct thump. Those parts of each measured room-average spectrum that lie between the strongly represented string-vibration components clearly show the envelope of the thump part of the net sound. As a matter of fact, the similarity between the shapes of the thump-envelopes of various notes, and of each one of these to the overall envelope displayed in Fig. 6, serves as a good confirmation of our picture of the piano sound-generation process.

Despite the fact that the piano sound is produced by an impulsive excitation taking place in only a very few milliseconds, the buildup of radiated sound in its neighborhood takes place over a period of time that is 10 to 50 times longer. We will begin our search for an explanation for this slow buildup by outlining the vibrational energy budget of a piano tone. Setting aside temporarily the energy associated with the initial thump, we can recognize that each string mode is abruptly supplied with its share of energy at the time of the hammer blow. Over the succeeding seconds, some of this energy is dissipated unproductively as heat within the body of the wire and at its anchorages. There is also a flow of vibrational energy into the soundboard, to set up its vibrations. Overall, the board vibrations build up under the stimulus of the string until the board's own energy loss rate to internal friction and to radiation into the room (and to some extent back into the string) are equal to the input rate from the string.

Globally speaking then, we would expect the radiated sound to rise in amplitude for a while (as the soundboard comes into equilibrium with the string vibrations) and then to decay gradually as the string gives up its energy. The initial part of a curve showing this behavior calculated for a typical pianolike string and soundboard is shown by the solid line in Fig. 7. This sort of calculation is able to give good account of the main feature of the onset times (35 to 50 ms). However, the measured behavior for a piano shows considerably more complexity, for the following reasons:

1. The initial thump is instantly transmitted to the soundboard, and the resulting wave travels across it and suffers numerous reflections.
2. During the initial epoch, while both the thump and the main tonal components are spreading across the soundboard, and making their first few reflections, all frequency components are able to radiate fairly efficiently. As a result, components in the sound output have a significant representation at very early times. This behavior is schematized by the irregular line in Fig. 7.
3. The cross-influences of the three piano strings and bridge belonging to a typical piano note make the strings' own decay quite irregular. This irregularity is then reflected in the longterm decay of the tone.

All this is an example of the statistical fluctuation behavior that was outlined for rooms in Section I of this article.

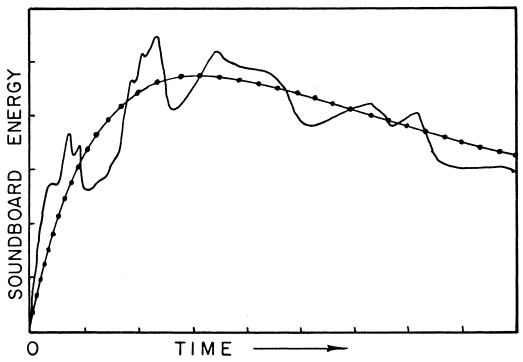


FIGURE 7 Smooth beaded curve: Trend of initial buildup of soundboard vibrational energy after excitation of a string. Irregular curve: Schematic representation of the actual buildup.

III. THE SINGING VOICE

The instruments discussed so far have belonged to a class of musical sound generators in which the primary source of acoustical energy (the vibrating string) is abruptly set in motion and allowed to die away over the next few seconds. There is another, much wider class of musical instruments (including the human voice, the woodwinds and brasses, and the violin family) in which the oscillations of the primary vibrator are able to sustain themselves over relatively long periods of time, drawing their energy from a nonoscillatory source such as the musician's wind supply or the steady motion of his bow arm.

The tone production system of the singing voice provides an excellent introduction to this class of continuous-tone instruments for two reasons. First, discussion is simplified by the fact that the primary exciter (the singer's larynx) maintains its own oscillations in a manner that is quasi-independent of the vocal tract air passages that it excites. Second, a further expository simplification comes about because the frequency of its oscillations is controlled by a set of muscles that are distinct from those that determine the shape of the upper airway.

Fundamentally, the larynx acts as a self-operating flow control valve that admits puffs of compressed air into the vocal tract in a regular sequence whose repetition rate f_0 determines the pitch of the note being sung. Since this flow signal is of a strictly periodic nature, the frequencies of its constituent sinusoids are exact integer multiples nf_0 of the laryngeal oscillation rate, and they therefore produce a sound of the type that is very well suited to the auditory processes of musical listening.

Aside from its ability to generate continuous rather than decaying sounds, the singer's voice-production mechanism, with its signal path from primary sound source (larynx) to concert hall via the vocal passages, is quite analogous in its physical behavior to that which leads the sound from vibrating string to the room by way of a soundboard or guitar body. Because the behavior of the sound transmission path through the vocal tract is far more important for present purposes than is the spectral description of the excitatory pulse train, we temporarily limit ourselves to the simple remark that the source component (of frequency nf_0) is related to the lowest component (of frequency f_0) by a factor $(1/n^2)A(n)$, where $A(n)$ has a relatively constant trend line plus a few irregularly spaced notches or quasi-notches. The perceptual significance of these notches is relatively small, as in the case of the stringed instrument spectra. In short, the spectrum envelope of any voiced sound in the room includes a factor $1/f^2$ due the source spectrum as a major contributor to its overall shape.

The vocal tract air column extending from the laryngeal source to the singer's open mouth may be analyzed as a nonuniform but essentially one-dimensional waveguide, whose detailed shape can be modified by actions of the throat, jaw, tongue, and lip muscles. One end of this duct is bounded by the high acoustical impedance presented by the larynx, and the other by the low impedance of the singer's open mouth aperture. Acoustical theory shows that such a bounded, one-dimensional medium has its natural frequencies spaced in a roughly uniform manner. Furthermore, the 15-cm length of this region implies that this mean spacing be about 1000 Hz, so that we expect no more than three or four such resonance frequencies in the region below 4000 Hz that contains the musically significant part of the voice spectrum. The signal transmission path from larynx to the listening room has a transfer function $T_r(f)$ that is the product of three factors. One is a term $T_1(f)$ falling smoothly as $1/f^{1/2}$ associated with acoustical energy losses that take place at the walls of the vocal tract. Another factor, $T_2(f)$, has to do with the efficacy of sound emission from the mouth aperture into the room. This rises smoothly with a magnitude proportional to the signal frequency f. The third factor, $T_3(f)$, fluctuates above and below a constant trend line, and it depends on the shape given to the vocal tract passage by its controlling muscles.

The peaks in $T_3(f)$ lie at frequencies that correspond to the normal-mode frequencies of the vocal tract if it is imagined to be closed off at the larynx and open at the mouth. The dips in the transmission function lie, on the other hand, at the modal frequencies of the vocal tract considered as an air column that is open at both ends. Both the peaks and dips have widths of about 50 Hz in the frequency range below 1500 Hz, rising to about 200 Hz at 4000 Hz. These peaks and dips tend to rise or fall above and below the trend line by about ± 10 dB (i.e., factors of about $3^{\pm 1}$). The nature of the overall vocal tract transfer function $T_{lr}(f)$ between larynx and the room is summarized as

$$T_{lr}(f) = (1/f)^{1/2} \times (f) \times (\text{peaks and dips})$$
$$= (f)^{1/2} \times (\text{peaks and dips}). \quad (10)$$

Because the vocal source spectrum has a relatively featureless $1/f^2$ behavior, it is convenient to display graphically the product $(1/f^2)T_{lr}(f)$, representing the sound normally measurable via the room-averaging procedure. Figure 8 presents such curves computed for three configurations of the vocal tract.

The pattern of peaks and dips in the $T_{lr}(f)$ function is of major perceptual significance: Each vowel or other voice sound is associated with a particular vocal tract configuration, and so a particular $T_{lr}(f)$. Speech then consists of a

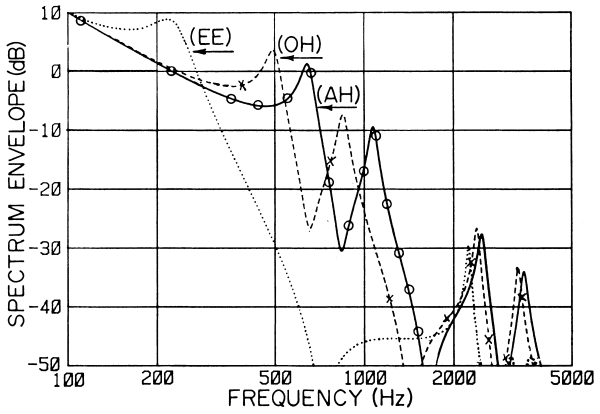


FIGURE 8 Schematic representation of spectrum envelope curves for three sung vowels.

rapidly changing set of spectral envelope patterns, which are recognized by the listener in a manner that depends almost not at all on the nature of the laryngeal source spectrum. Thus, a singer who produce the vowel aah at a pitch of A_2 ($f_0 = 110$ Hz) supplies his listeners with a generated sound whose harmonic components can be evaluated from the curve for aah (see Fig. 8) at the discrete frequencies 110, 220, 330, ... Hz, as indicated by the small circles on the curve. Similarly, a singer producing the vowel ooh at G_4 ($f_0 = 392$ Hz) emits a sound whose spectrum has component amplitudes that are related in the manner indicated by the small x's.

The vowel pattern recognition abilities of the human listener are highly developed. Whispered speech is perfectly comprehensible, even though the source signal consists of a densely distributed random collection of sinusoids (white noise) rather than the discrete and harmonic collection of voiced speech. Furthermore, a radio announcer is completely intelligible whether the receiver tone controls are set to "treble boost, bass cut" (nearly equivalent to multiplying $T_{lr}(f)$ by f) or to "treble cut, bass boost" (which is nearly the same as multiplying the spectrum by $1/f$). The recognition process requires, as a matter of fact, only the existence of properly located peaks relative to the local trend of the spectrum, while the positions and depths of the dips are essentially irrelevant, to the point where many electronic speech synthesizers omit them entirely. Thus, all that is really necessary is to specify the frequencies of the lowest three or four transmission peaks for each sound. These are denoted by F_1, F_2, F_3, and F_4 and are called the format frequencies (and are about 17% higher for women than for men).

So far, no clear distinction needs to be made between speech and song beyond the need in music for precisely defined pitch (and thence values of f_0). The musician has, however, three special resources that have little significance in speech. First of all, the source spectrum shape, which is approximated by $A_n = A_1/n^2$ for the "mildest" and most speechlike tone color, may be modified to the form

$$A_n = A_1\left[\frac{1+(1/\gamma)^2}{1+(n/\gamma)^2}\right]. \tag{11}$$

The components for which n is less than γ have then essentially the same amplitude as A_1, while the $1/n^2$ fall-off is postponed to the higher, $h \gg \delta$ components. In the extreme case, γ can be as large as 3.

The second resource of the singer is the use of what is known as format tuning. Consider a soprano who is asked to sing the vowel aah at the pitch D_5 at the top of the treble staff. This means that her sound consists of sinusoids having frequencies close to $nf_0 = 587, 1175, 1762, \ldots$ Hz. Her normally spoken "aah" would have a second-formant frequency F_2, close to 1260 Hz, but she may choose to alter her vocal tract shape (and thus modify the vowel sound) somewhat, in order to place F_2 exactly on top of the 1175-Hz second harmonic of the sung note. This sort of tuning is effective only for notes sung at a pitch high enough that a formant frequency can be adjusted to one of the first three voice harmonics, which assures that the rapid $1/n^{3/2}$ fall-off in amplitude has not significantly reduced the prominence of the tuned component in the net sound. The most obvious benefit to be gained from formant tuning is almost trivial: The net loudness of the note is increased to an extent that can be useful if the singer is struggling for audibility against an overpowering accompanist. Subtler, and more significant musically, is a sort of glow and fullness that is imparted to the tone of a formant-tuned note. The perceptual reasons for this are not entirely clear, but the fame of many fine sopranos is enhanced by their skillful use of the technique.

The third spectral modification that is available to singers (especially for tenors, and for the highest notes of other males) is a rather curious one: A systematic modification of the vocal tract region near the larynx and/or a manner of vowel production that makes the second and third formant frequencies almost coincident can give rise to an extremely strong transmission peak in the neighborhood of 3000 Hz. This peak is referred to as the singer's formant, regardless of its mode of production. The presence of such a formant considerably increases the net sound output power of the voice, a fact that joins with certain features of the ear's perception mechanism to produce a large increase in the loudness of all tones sung in this way. It also produces what is usually referred to as tonal brilliance and penetrating character.

When used flexibly and tastefully by true artists all three of these vocal resources greatly enhance the beauty and

expressiveness of the musical line. For them to have such effects, they must be used subtly, with close attention to the meaning of the music and the words. As a group, lesser singers do not make use of formant tuning except to increase their loudness. Among this same group of less-than-satisfactory performers, the other two forms of vocal production are used incessantly, in part to call attention to themselves by mere loudness, and in part as evidence that they have what is called a "trained voice." It is a curious fact that if any other musical instrument (or a loudspeaker) had a strong and invariable peak around 3000 Hz in its spectral envelope, it would be subject to instant and bitter criticism.

IV. THE WIND INSTRUMENTS

The first and most important feature of the family of wind instruments (from the point of view of physics and perception psychology) is the fact that its tones are self-sustaining. The duration of these tones is limited only by the desire of the player and the sufficiency of his air supply. As hinted already in connection with the oscillations in the singer's larynx, self-sustaining oscillators of necessity give rise to sounds made up of exactly harmonic components.

The second distinguishing feature of the wind instruments is that the air column whose natural frequencies control the frequency and wave shape of the primary oscillation is also the device that transmits the resulting sounds to the listening room. It is no longer possible (as with the stringed instruments and with the voice) to describe a vibration source that is essentially independent of the transmission mechanisms that convert its output into the sounds that we hear.

A. The Structure of a Wind Instrument

Figure 9 will serve to introduce the essential features of a musical wind instrument as it is seen by a physicist.

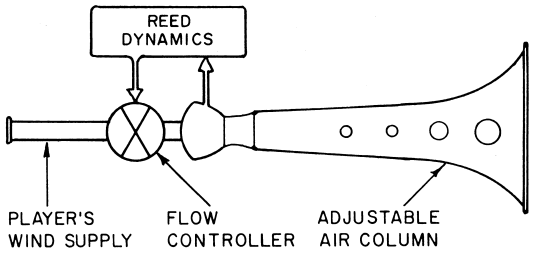


FIGURE 9 Basic structure of a wind instrument: Air supply, flow controller, air column, and dynamical coupling between the latter two.

To begin with, the player is responsible for providing a supply of compressed air to the instrument's reed system. This reed system functions as a flow controller that admits puffs of air into an adjustable air column belonging to the instrument itself. The system oscillates because the flow controller is actuated by an acoustical signal generated within the upper end of the air column; this signal is in fact the air column's response to the excitory flow injected via the reed.

The structural features that serve to distinguish between the two major families of wind instruments may be summarized as follows:

1. A woodwind is recognized by the fact that the length of its air column is adjusted by means of a sequence of toneholes that are opened or closed in various combinations to determine the desired notes. The oboe, clarinet, saxophone, and flute are all members of this family.
2. A brass instrument is distinguished by the fact that its air column continues uninterrupted from mouthpiece to bell, the necessary length adjustments being provided either via segments of additional tubing that are added into the bore by means of valves as in the trumpet or by means of a sliding extension of the sort found on the trombone.

The sound production process in all wind instruments involves the action of an air flow controller under the influence of acoustical disturbances produced within the air column. This provides another description of the various kinds of wind instrument in terms of the flow controllers that are found on the different instruments.

1. The cane reed is found on the clarinet, oboe, bassoon, and saxophone. It is not (for present purposes) necessary to further distinguish between single and double reeds; they both share the dynamical property that the valve action is such as to decrease the air flow through them when the pressure is increased within the player's mouth.
2. The lip reed normally used on brass instruments and the cornetto is the second major type of flow controller. Here the valve action is such that the transmitted flow is increased by an increment of the pressure in the player's mouth.
3. Flutes, recorders, and most organ pipes are kept in oscillation through the action of a third type of controller that may aptly be described as an air reed. Here we find an air jet whose path is deflected into and out of the air column through the action of the velocity of the air as it oscillates up and down the length of the governing air column.

It should be emphasized that while the nature of the flow controller itself is very important, it does not usefully

distinguish the instrumental wind families from each other in the essential features of their oscillatory behavior.

B. The Oscillation Process: Time-Domain Version

There are two main ways of describing the oscillation processes of a self-sustained instrument. The time-domain description deals with the temporal growth and evolution of a small initial impulse that leaves the flow controller and then is reflected back and forth between the two terminations of the air column—the tone holes and/or bell at its lower-end termination and the flow-controlling reed system at its upper end.

When the initial impulse is reflected from the lower termination, it suffers a change of form and an enfeeblement of its vigor (as does each of its successors). The change in form arises because of the acoustical complexity of the termination, and the loss in amplitude occurs because some of the incident wave energy has been lost in the journey and some transmitted into the outside air. At the reed end, another very different form of terminating complexity produces a change in the reflected shape of each returning impulse that travels up to it from the bottom end of the instrument. The size of this regenerated pulse also increases because it receives energy supplied by the incoming compressed air from the player's lungs.

As the tonal start-up process evolves toward the condition of steady oscillation, the wave shape stabilizes into one in which each reflection at the lower end of the air column is modified in such a way as to "undo" the modification that takes place at the reed.

C. The Frequency-Domain Description of Wind Instrument Oscillation

The time-domain description of the oscillation process is readily susceptible to mathematical analysis and permits detailed calculation of the sound spectra produced by a given reed and air column. However, it is ill-adapted to the task of showing general relations between the mechanical structure of an instrument and its playing behavior, nor will it guide its maker in adjusting it for improved tone and response.

Fortunately, a second way of picturing the oscillatory system—the frequency-domain version—can readily deal with such questions and is well suited for our present descriptive purposes. In the frequency-domain analysis, we start by relating the proportions of an instrument to the natural frequencies and dampings of the various vibrational modes of the controlling air column. For present purposes, it suffices to describe the flow controller merely by reiterating that the increment of flow produced by an increment of control signal is not in proportion to it. In particular, a sinusoidal control signal of frequency $f = P$ gives rise to a pulsating flow that may be analyzed into constituent sinusoids having a harmonic set of frequencies $P, 2P, 3P, \ldots$. Additional components appear when the excitatory signal is itself the superposition of several sinusoids. If these have the frequencies $P, Q, R, S, \ldots$, the resulting flow signal will contain an elaborate collection of components having frequencies that can be described by

$$f = |\alpha P \pm \beta Q \pm \gamma R \pm \delta S|. \tag{12}$$

Here α, β, γ, and δ are integers that can take on any values between zero and an upper limit N, which can be as high as 4 or 5. Clearly, hundreds of these frequencies can be present in the flow. (Because of their cross-bred ancestry, they are known as heterodyne frequencies.) It is also clear from their very number and their computational origins that they are distributed over a frequency range extending from zero to more than N times the highest of the stimulus frequencies, and that the amplitude of each of these new components is determined by a combination of the amplitudes of all the original components. This means that the energy associated with each flow component is determined jointly by all members of the controlling set of sinusoids. It is this cross-coupling of stimuli and responses having widely different frequencies that underlies the dynamical behavior of all self-sustained musical instruments, and so governs their musical properties.

Consider the behavior of a reed coupled to an air column designed in such a way as to have only a single resonant mode of oscillation. When blown softly, the system will oscillate at a frequency f_0 that is very nearly equal to the modal frequency. Because the strengths of the higher numbered heterodyne components always fall toward zero under conditions of weak excitation, almost the entire efficacy of the flow controller is focused on supplying excitation to the air column at the frequency f_0 of its own maximum response, and the system can oscillate efficiently. However, the system is perfectly stable, because any tendency of the system to "run away" leads to the production of heterodyne components that dissipate energy. These components do not replenish themselves because the air column does not respond strongly to them and so does not "instruct" the reed to reproduce them. If we try to play loudly by blowing harder on such a single-mode instrument, the sinusoid at f_0 hardly changes in strength, but some hissing noise appears and the reed either chokes up entirely (on the cane reed instruments) or blows wide open (on the brasses), with a complete cessation of tone.

Similar behavior is observed for a multimode air column if the mode frequencies are randomly placed. Usually the system starts as though only the strongest resonance was present, and the choking-up of the oscillation is more

abrupt because of the enormously increased number of unproductive heterodyne components that are produced when the blowing pressure is increased.

The discussion so far has shown the conditions under which a reed-plus-air-column system can not play; it is time now to describe the requirements for a system that will produce sounds other than weak sinusoids. Suppose that the air column has a shape such that its natural frequencies themselves form a (very nearly) harmonic set, in the manner

$$f_n = nf_1 + \varepsilon_n, \tag{13}$$

where the discrepancy ε_n is a measure of the inharmonicity. The heterodyne frequencies will now form small clumps closely grouped around the exactly harmonic frequencies nf_n. As a group, the modes thus appear able then to cooperate with the reed and so to regenerate the flow-stimulus energy (distributed in narrow clumps of ever-growing complexity). What actually happens is that the modes quickly lock together to produce a strictly harmonic oscillation with a repetition rate f_0 such that the overall energy production of the system is maximized. Such a mode-locked regime of oscillation turns out to be increasingly quick-starting and stable in all respects if the air-column mode frequencies are increasingly well aligned into harmonic relationship. It will also run over a wide range of blowing pressures (and so produce a musically useful range of loudness).

D. Musically Useful Air-Column Shapes

We have just learned that for a self-sustained multimode oscillation to exist at all, the air-column shape must be such that the natural frequencies of its modes are in very nearly exact harmonic relationships. There are very few possible basic shapes that can meet this criterion. For instruments of the reed woodwind type there are two, for the brass instruments there are two, and for the air-reed (flute) family there is only one. It can be shown that because the cane-reed and lip-reed instruments have pressure-operated flow controllers, the relevant air column's natural frequencies are those calculated or measured under the condition that its blowing end be closed off by means of an air-tight plug, while the downstream end is left in open communication with the outside air via the tone holes and bell. On the other hand, for the velocity-operated air reed of the flute family, it is necessary to consider the air-column modal frequencies for the condition when both ends are open.

The clarinet family is the sole representative of the cylindrical bore (first) type of possible reed woodwind while the oboe, bassoon, and saxophone belong to the basically conical second group. The trumpet, trombone, and French horn are representatives of the outward-flaring hyperbolic-shaped air columns suitable for brass instruments, while the flügelhorn and certain baritone horns are familiar examples of the conical second group. The flutes, on the other hand, are all based on a straight-sided tube, which can have positive, negative, or zero taper. That is, they can either expand or contract conically in going downstream from the blowing end, or can be untapering (i.e., cylindrical).

Because of the acoustical complexities of the tone holes at the lower ends of all the woodwinds and of the mouthpiece and reed structures at the upper ends of both brasses and woodwinds, the actual air-column shapes of the various instruments differ in many small ways from their prototypical bases. In all cases, however the differences are such as to align the modal frequencies of the complete air column in the required harmonic relationship.

E. Sound Spectra in the Mouthpiece/Reed Cavity

It would require a lengthy discussion to explain the ways in which the mouthpiece/reed cavity sound pressure spectrum (which "instructs" the flow-controlling reed) takes its form, but it is not difficult to describe its general nature for the various kinds of wind instrument.

It is clear that a lot of the f_0 fundamental component will be present in the mouthpiece spectrum: not only is it directly generated via the lowest frequency air-column resonance but also by difference-heterodyne action between every adjacent pair of the higher harmonics. In similar fashion, there will be a fair amount of heterodyne contribution to the second harmonic component arising from (at the very least) the nonlinear interaction of every alternate pair of harmonics. Analogous contributions are likewise made to the higher tonal components by ever more complex combinations of pairs of peaks in the resonance curve.

Details aside, the foregoing considerations are themselves able to imply a tendency for the successive harmonics of the generated tone to be progressively weaker. A closer look brings in information about the manner in which the properties of the air column and reed act to determine the spectrum. For a cane-reed woodwind (such as a clarinet, saxophone, or oboe) or for a lip-reed brass instrument it is possible to show that as long as the playing level is low enough that the reed does not pound completely closed during any part of its swing, the behavior of the pressure amplitudes p_n of the various harmonics is well caricatured by

$$p_n = \left(\frac{p_1}{p_0}\right)^n \left\{\frac{Z_n F(f_n) + M_n}{[F(f_n) - Z_n] + D_n}\right\}. \tag{14}$$

Here Z_n is the height of the flow-induced resonance response curve (input impedance) at the frequency f_n of the

nth harmonic component of the played tone; the factor $(p_i/p_0)^n$ gives the influence of the playing level on the overall spectrum; p_1 is the amplitude of the fundamental component of the tone; and p_0 is a reference pressure defined such that $(p_1/p_0) = 1$ when the reed just closes at one extreme of its cyclic swing. We shall postpone an explanation of the musical recognizability of p_0 or its relation to the ordinary loudness specifications that run from pianissimo to fortissimo in the player's natural vocabulary and will note only that an important measure of the strength of any component is the magnitude of its associated air-column resonance peak.

The functions M_n and D_n are slowly changing, and they describe the already-mentioned nonlinear processes of energy exchange between spectral components that assure amplitude stability and well-defined waveforms at all playing levels. The fuction $F(f_n)$ describes the relation between the reed's primary flow-controlling ability and the frequency of some signal component that may be acting upon it:

$$F(f_n) = \pm K_r[1 - (f_n/f_r)^2]. \tag{15}$$

Here K_r is a constant, f_n is the frequency of the signal component, and f_r is the (player-controllable) natural frequency of the reed taken by itself. The plus sign in the defining equation for F applies to the cane reeds (which are pushed open by an increase in mouthpiece pressure), while the minus sign applies to the lip reeds belonging to the ordinary brass instruments (which are pushed closed by an increase in mouthpiece pressure). Since F must be positive if oscillation is to be supported, the lowest few harmonics of the tone of cane reed instruments must have $f_n \leq f_r$, whereas for the brasses $f_n \geq f_r$ for all the components.

Figure 10 illustrates the state of affairs for the woodwind family of instruments. The measured input impedance curve $Z(f)$ is shown for an English horn fingered to give

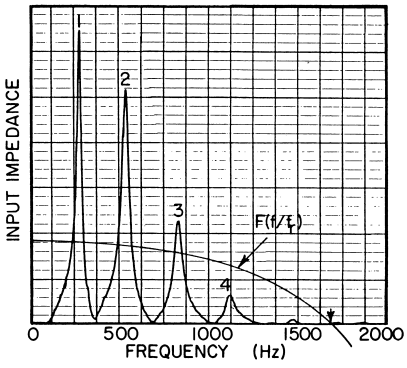


FIGURE 10 Measured resonance curve for a typical woodwind air column (English horn), along with a curve showing the nature of the primary flow-control function $F(f)$.

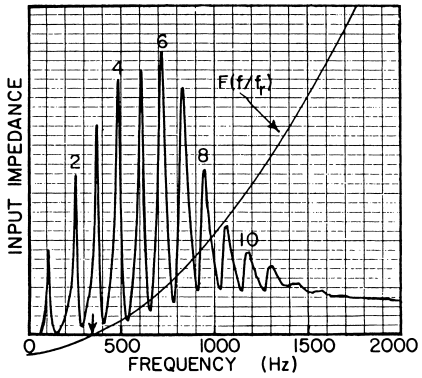


FIGURE 11 Measured resonance curve for a brass instrument air column (trumpet). It is shown along with a typical brass instrument flow-control function $F(f)$.

the air column for playing its own (written) note C_4. Also shown is a typical flow-control curve $F(f)$ with the reed frequency set at 1650 Hz. Notice (for future reference) that the air column almost completely lacks resonance peaks above what is known as its cutoff frequency f_c, which lies near 1200 Hz.

Figure 11 shows in a similar fashion the impedance curve and $F(f)$ for a trumpet in the case in which the player has set $f_r \simeq 350$ Hz in preparation for sounding his written note G_4. This note has its major energy production associated with the cooperation of resonance peaks 3, 6, and 9 (which are in accurately harmonic relationship on a good trumpet). Once again we call attention to the absence of air-column resonances above a cutoff frequency, this time lying near 1500 Hz. The pressure spectrum in the trumpet's mouthpiece is difficult to guess by eye because it depends on the product of the heights of the Z_n peaks and the rising $F(f)$ curve; however, it is clear that the components having frequencies above the 1500-Hz cutoff are very weak. There are thus two reasons (acting for both woodwinds and brasses) why the spectrum should have a strong first harmonic and progressively weaker second, third, and fourth components, with rapidly disappearing components above that. In addition to the falloff associated with the weakening heterodyne contribution, we have at high frequencies a progressive reduction in the resonance peak heights, and above f_c all energy production ceases.

So far, resonance curves have been presented for only one of the many air-column configurations that are possible via the fingerings available to a player. It goes almost without saying that when a player desires to sound a lower note, he lengthens the air column of a woodwind by closing a tonehole, or by adding an extra length of tubing to the bore of a brass instrument by means of a valve piston or slide. In this way the frequencies of all the resonance peaks are shifted downward by a factor of 1/1.05946 for every semitone lowering of the desired pitch. An example

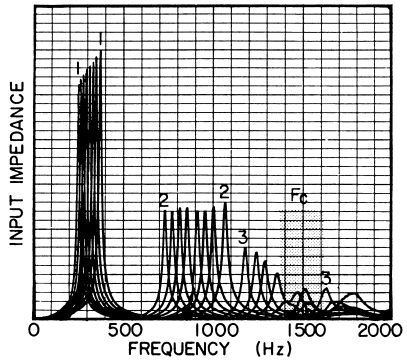


FIGURE 12 Clarinet resonance curves measured for the air columns used in playing the notes C_4 to G_4.

of this behavior is presented in Fig. 12, where the resonance curves are presented for the written notes lying between C_4 and G_4 of a clarinet. The leftmost peak labeled by the numeral 1 is the first-mode resonance for the air column used to produce C_4; the leftmost peak marked with a 2 similarly indicates the second-mode peak belonging to the same column arrangement, and so on for the higher-numbered peaks. In an exactly parallel way, the rightmost numerals 1, 2, 3, . . . indicate the corresponding resonance peaks for the note G_4. There are two noteworthy features in this set of impedance curves. The first is shared by all wind instrument resonance curves: The cutoff frequency (above which there are no peaks) remains the same for all fingerings. The second feature is characteristic of the clarinet family alone: While peaks 1, 2, 3, . . . are in the strict whole-number relationship demanded by the cooperative nature of wind instrument tone production, the modal frequencies f_1, f_2, f_3, . . . lie in a 1, 3, 5, . . . sequence, with dips in the resonance curves appearing at the positions of the even multiples of the mode-1 frequency. An immediate consequence of this fact is that despite the restorative powers of the alternate-component heterodynes, the even-numbered members of the generated mouth-piece pressure spectrum are weaker than the odd-numbered ones. Because of the fixed cutoff frequency shared by all the air columns used to play notes on a given instrument, and because all the notes share the same mouthpiece and reed structure, it is possible to construct spectrum envelope formulas for the notes of the various classes of instrument. These formulas have a mathematical structure very much like those presented earlier in connection with the bridge-driving forces exerted by the strings of a guitar, harpsichord, or piano. The basic physics that determines them is of course entirely different here, since wind instrument oscillations are active, nonlinear, and self-sustaining.

It is fairly obvious from the mathematical nature of the $(p_i/p_0)^n$ factor that if the player sounds his notes progressively more softly (without changing the tension of his lips or the setting of the reed), the higher-n components fall away very much more quickly than the lower members of the sequence. In decibel language, we can say that for every decibel that the p_1 component is weakened, the level of p_n falls by n dB. At the softest possible level, then, we expect the tone within the mouthpiece to have degenerated into a single sinusoid of frequency f_1.

In actual practice, the oboe is essentially unplayable at levels for which $(p_i/p_0) < 1$, while the bassoon is almost never played thus. The saxophone can be so used, but normally it too is used in the domain where $p_i/p_0 > 1$. Only the clarinet is played at the levels discussed so far, and for it, the customary forte instruction gives a tone for which p_i/p_0 is little or no larger than unity. This raises the immediate question of what happens to the spectral envelope when an instrument is played at the higher dynamic levels. The answer varies with the instrument, and while it is fairly well known, it would take us too far afield to discuss it here. The solid curve in Fig. 13 presents the general shape of the *internal* (mouthpiece) spectral envelope for the nonclarinet woodwinds. The corresponding internal spectral envelope for the brasses is shown in the same figure by a closely dotted curve, while the behavior of the odd and even components of the clarinet's spectrum is shown by the pair of dashed lines. All these curves are calculated on the assumption that the factor $(p_1/p_0)^n$ belongs to the instrument's normal mezzoforte playing level.

The interplay between the direct energy processes at the air-column resonance peaks and the heterodyne transfer of energy between components is made vivid by the following observations. The heights of the various resonance peaks in Fig. 11 show clearly that there can be little direct

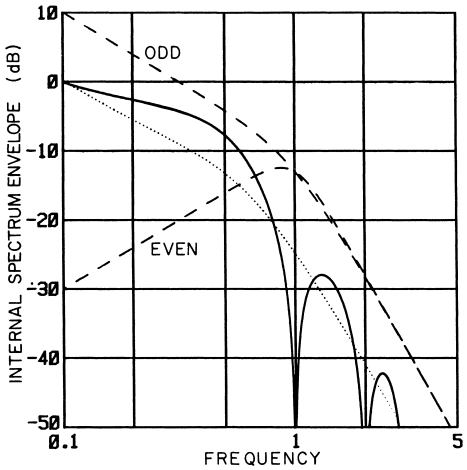


FIGURE 13 Internal (mouthpiece or reed-cavity) spectrum envelopes. Nonclarinet woodwinds, solid curve; brasses, dotted curve; clarinet, dashed curves, one for the odd-numbered components of the played note, and one for the evens.

production of energy by the first two or three components of the played tone C_4 (led cooperatively by peaks 2, 4, 6, . . .).

For brass instruments, on the other hand, the dotted spectrum envelope curve given in Fig. 13 shows, that the actual strength of the generated fundamental component in the mouthpiece is the strongest of all, while components 2, 3, and 4 are progressively weaker. Clearly, a great deal of heterodyne action is needed to transfer the majority of the high-frequency generated power into the low-frequency part of the spectrum. We shall meet similar behavior in the tone production processes of the violin.

Consider next what happens when the trumpet player sounds what he calls the F_2 pedal note, whose repetition rate is exactly one-third that of C_4. The harmonic components 3, 6, 9, . . . of the pedal tone lie at air-column peaks 2, 4, 6, . . . and so can act as direct producers of acoustical energy. Meanwhile, the tonal components (1, 2), (4, 5), (7, 8), . . . are away from any resonance peaks, and so exist only because of the heterodyne conversion process. The shape of the measured mouthpiece spectrum envelope for components 3, 6, 9 . . . for F_2 is almost identical with that belonging to C_4. The remaining (heterodyned) components have a very similar envelope, but one that is many decibels weaker.

F. Transformation of the Mouthpiece Spectrum Envelope into the Room-Average Envelope

Attention has already been called to the fact that regardless of the air-column configuration, each musical wind instrument has a cutoff frequency f_c above which it lacks resonance response peaks. The same air-column physics that produces a falling away of the resonance peak heights immediately below f_c (and so a reduced production of the corresponding mouthpiece pressure components) also plays a significant role in the transformation of the mouthpiece sound into the one enjoyed by listeners in the concert hall.

It is a property of the sequence of open tone holes at the lower end of a woodwind that, at low frequencies, sound is almost exclusively radiated from only the first of the holes, while the lower holes come into active play one by one as the signal frequency is raised. Above the cutoff frequency f_c, all of the holes are fully active as radiators, and the sound emission not only becomes nearly independent of frequency, but also essentially complete.

While the acoustical laws of sound transmission and radiation from a brass instrument bell are quite different from those governing the woodwind tone holes, here too we find very weak emission of low-frequency sounds. The bells's radiation effectiveness then rises steadily with frequency until it is once more complete for signals above f_c.

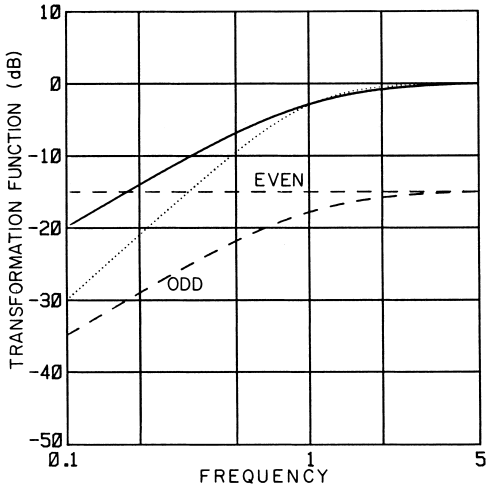


FIGURE 14 Spectrum transformation function converting the internal spectrum envelope into the room-average one. Nonclarinet woodwinds, solid curve; brasses, dotted line; clarinet, a generally rising dashed line for the odd-numbered tonal components and a horizontal dashed line for the even components. For clarity both clarinet curves have been displaced downward 15 dB from the other curves.

All this explains why there are no resonance peaks above f_c, and why those having frequencies just below f_c are not very high: the energy loss associated with radiation acts to provide a frequency-dependent damping on the resonance, a damping that becomes complete above f_c. In other words, the same phenomena that increase the emission of high-frequency sound from the interior of a wind instrument to the room around it also lead to a progressively falling ability of the instrument to generate high-frequency sounds within itself (this is shown by the dotted curve in Fig. 13).

The solid line in Fig. 14 shows the behavior of nonclarinet woodwinds in transferring their internal sounds into the room. Note that the transfer becomes essentially complete for components having frequencies above the cutoff frequency. The dotted and the dashed curves of this figure show the analogous spectrum transformation function for the brass instruments and for the odd and even components of the clarinet spectrum.

G. Overall Spectrum Envelopes of Wind Instruments in a Room

Figure 15 illustrates the nature of the room-average sound-spectrum envelopes of the main reed instrument classes, as calculated from the curves for their mouthpiece spectra and their transformation functions. Once again the solid curve pertains to the nonclarinet woodwinds, the dotted line to the brasses, and the pair of dashed line to the clarinets. The essential correctness of these diagrammatic

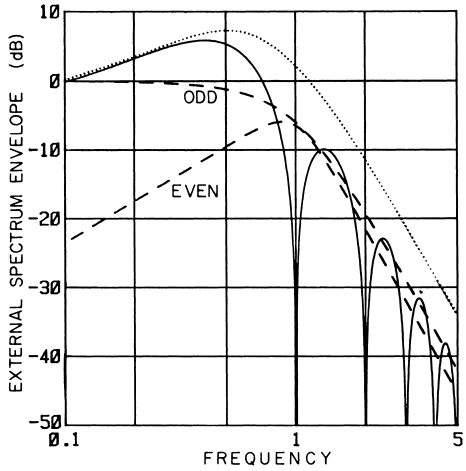


FIGURE 15 External spectrum envelopes. Nonclarinet woodwinds, solid line; brasses, dotted line; clarinet, a pair of dashed lines: one for the odd and one for the even components of the tone.

representations is clearly shown in Fig. 16a–c, which presents the room-average spectrum envelopes measured for the oboe (C_4 to C_6), trumpet (E_3 to $F_5^{\#}$), and clarinet (E_3 to C_6). These were obtained using room-averaging techniques quite similar to those used to obtain the spectrum envelope of a piano (see Fig. 6).

Exactly as is the case for the purely heterodyned even-harmonic components of the clarinet tone, the (1, 2), (4, 5) (7, 8), ... components of the trumpet F_2 pedal note are weakly generated but strongly radiated. As a result, in the measured room-averaged spectrum of this note these components essentially fit the envelope belonging to the directly generated 3, 6, 9 . . . components (being only about 3 dB weaker).

Study of a large variety of instruments (including soprano, alto, and bass representatives of each family) shows that the basic curves change very little from one example to another. In all cases the general trend at high frequencies is for the envelope to fall away as $1/f^3$, with a breakpoint close to 1500 Hz for the soprano instruments (oboe, trumpet, clarinet). For the alto instruments (English horn, alto saxophone, alto clarinet), the breakpoint is around 1000 Hz, paralleling the fact that their playing range lies a musical fifth below the soprano instruments. For the next lower range of instruments (trombone, tenor saxophone, bass clarinet), the break lies around $1500/2 = 750$ Hz, while the bassoon (whose playing range lies at one-third the frequency of the oboe) has its break near $1500/3 = 500$ Hz.

Almost nothing has been said so far about the flute family of woodwinds beyond a description of its associated flow controller and the basic nature of its usable air column. Despite the flute's apparent mechanical simplicity, it is in many ways dynamically more subtle than the other woodwinds and somewhat less well understood.

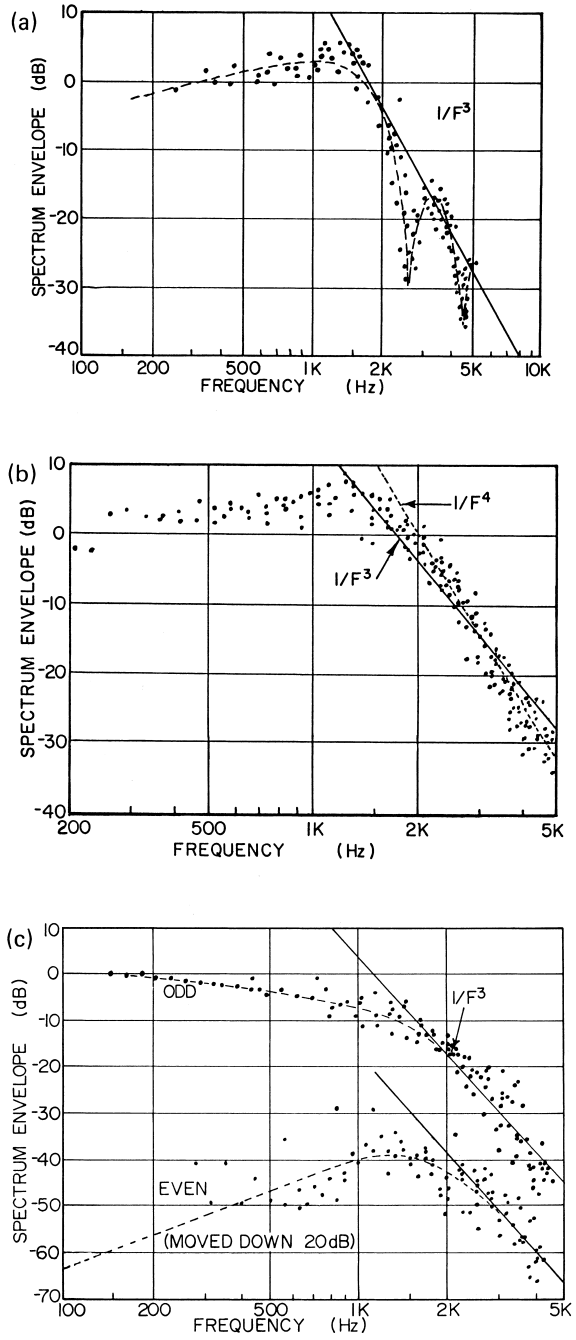


FIGURE 16 Measured room-average spectra (a) oboe; (b) trumpet; (c) clarinet, for clarity the even-component data have been displaced downward by 20 dB.

Because the flute's flow controller is velocity operated rather than pressure operated, a somewhat roundabout proof shows that it is the peaks in the admittance curve (flow/pressure) rather than those of its reciprocal, the impedance curve, that cooperate with and instruct the

flute's air reed. The room-average spectrum of a flute may be expected a priori to be quite different from that of the other woodwinds for two reasons. First, the oscillation dynamics of the primary energy production mechanism are drastically different. Second, there are two sources of sound radiation into the room: one is the familiar one associated with the tone hole lattice ($f_c \simeq 2000$ Hz), while the other is the oscillatory flow at the embouchure hole across which the player blows. It is somewhat as though the room were supplied simultaneously with the internal and the external spectrum of a normal woodwind!

The room-average spectrum envelope for a flute has a curiously simple form that is well represented by

$$E(f) = e^{-f/f_a}. \tag{16}$$

Here f_a is near 800 Hz for the ordinary concert flute, close to 530 Hz for the alto, and 1600 Hz for the piccolo (as might be expected for such systematically scaled instruments).

V. THE BOWED STRING INSTRUMENTS

The bowed string family of musical instruments shares many features with the instruments that have been discussed so far. For this reason the present section can serve both as a review and elaboration of the earlier material and as an introduction to another major class of instruments.

As indicated in Fig. 17, the violin, like the guitar, has a boxlike structure with a rigid neck and a set of strings whose vibrating length can be controlled by the player's fingers. To a first approximation then, the two instrumental types have similar dynamical processes that convert the driving-force spectrum (transferred from the strings via the bridge) to the sound spectrum that is measured in the concert hall. On the other hand, the excitation mechanism of the self-sustained string oscillation of violin family instruments proves to be essentially the same as that which generates sound in the woodwind and brass instruments.

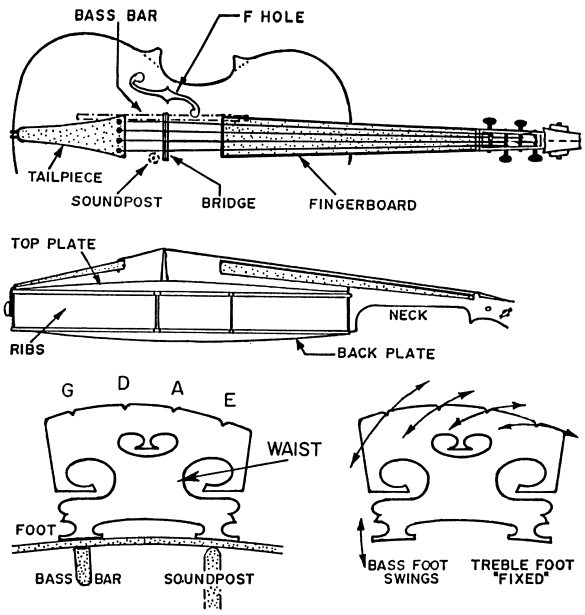


FIGURE 17 Structural parts of the violin, and their names.

A. The Excitatory Functioning of the Bow

The mechanism used by the violin family to keep a string in vibration is easily sketched: The frictional force exerted at the contact point between bow and string is smaller when there is a fast slipping of the bow hair over the string and larger when the sliding rate is slower. Thus, during those parts of its oscillation cycle when the contact point of the string chances to be swinging in the same direction as that of the rapidly moving bow (so that the slipping velocity is small), there is a strong frictional force urging the string forward in the direction of its motion. During the other half of the cycle, the string is moving in a direction opposite to that of the bow. Under these conditions the slipping velocity is large, making the frictional force quite small. Notice that this frictional drag is still exerted in the direction of the (forward) bow motion, and it therefore acts to somewhat retard the (backward) vibrational motion of the string. In short, during part of each vibratory cycle a strong force acts to augment the oscillation, and during the remainder of the cycle there is a weaker depleting action. The oscillation builds up until the bow's energy augmentation process exactly offsets all forms of energy dissipation that may take place at the bowing point and everywhere else. It is clear that the vigor of the oscillation is ultimately limited by the fact that during its "forward" swing the string velocity at the bowing point can equal, but not exceed, the velocity of the bow, otherwise the friction would reverse itself and pull the string velocity back down to match that of the bow.

In the earliest version of the formal theory of the bowed string excitation process, it was assumed that the string and bow hair stick and move together during the forward motion of the string, while during the return swing the friction is negligible. Such a theory is remarkably successful in predicting many of the most obvious features of the oscillation but is powerles to give a dependable account of the actual driving-force spectrum envelope as it might appear at the bridge.

B. The Frequency-Domain Formulation Reapplied

In the framework of the resonance-curve/excitation-controller theory of self-sustained oscillators, it is easy

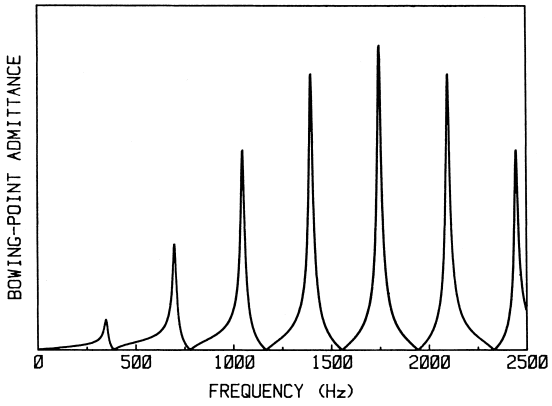


FIGURE 18 Violin bowing-point admittance (velocity/force) curve for F f
the way along the string from the bridge to the player's left-hand fingers.

to see that the bow/string interaction serves as a velocity-operated force controller; this is in contrast to the pressure-operated flow controllers found in the woodwinds and brasses. To be consistent then, in our theory we replac the pressure-response curves of flow-driven air columns (as measured in the mouthpiece) by the velocity-response curve of a force-driven string (measured at the bowing point). Figure 18 shows a velocity-response (driving-point admittance) curve calculated for a violin D string fingered to play F_4. In this example, it has been assumed that the bow crosses the string at a point one-tenth of the string length away from the bridge (about 25 mm). Notice the remarkable similarity of this resonance curve to the one shown in Fig. 11 for the air column of a trumpet.

For the trumpet, the increasing height of the resonance peaks and the initial falling away beyond their maximum is determined chiefly by the design of the mouthpiece cup (a cavity) and back bore (a constriction). The ultimate disappearance of the peaks is, as already explained, due to the radiation loss to the room suffered by the air column. In the case of a violin however, it is the distance of the bowing point from the bridge that determines the frequency region (around 1750 Hz in the example) where the admittance peaks are tallest. The subsequent weakening of the higher frequency peaks is controlled jointly by the rising influence of frictional and radiation damping and by some bow physics that is related to that which produces "notches" in the plucked string's $E_a(f)$ function [see Eq. (5)].

There exists a force-control function representing the bow/string interaction that is analogous to the wind instrument $F(f)$ function [see Eq. (15)]. While this analog to F is not shown in Fig. 18, it may be taken to be quite similar to the one shown for woodwinds in Fig. 10. There is unfortunately no simple description for the bow properties that together play the role of the reed resonance frequency f_r in limiting the production of energy at high frequencies.

C. Spectrum Systematics

The small height of the lowest few resonance peaks in Fig. 18 shows that the major part of the total energy production comes via the tall response peaks that lie at higher frequencies. Not surprisingly, the nonlinear nature of the bow/string stick-slip force results in heterodyne effects that lead to a bowing point spectrum very similar to that of a trumpet mouthpiece pressure spectrum. The systematic transfer of high-frequency energy into low-frequency vibrational components is as effective for the violin as for the trumpet, so that (as in the trumpet) the driving-point spectrum ends up with the first component strongest and higher ones becoming progressively weaker.

The simplest stick-slip theory of the bow/string oscillation gives a reasonably accurate initial picture of the spectrum envelope $E_v(f)$ for the string velocity at the bowing point:

$$E_v(f) = [\sin(\pi f/f_\beta)]/(\pi f/f_\beta). \tag{17}$$

Here f_β is defined in terms of the point of application of the bow on the string in exactly the same manner as f_a was defined via the plucking point in Eq. (5). The shape of $E_v(f)$ is shown exactly by the curve for $E_a(f)$ in Fig. 4. The actual velocity spectrum envelope of a bowed string is quite similar to that implied by Eq. (17), except that (a) the spectral notches do not go all the way to zero, and (b) at high frequencies the effects of dampings, etc., reduce the spectral amplitudes very considerably.

So far the discussion of the spectral properties of the string velocity at the bowing point serves as a means for clarifying the fundamental energy production processes of the bowed string. What actually leads to the radiation of sound in the room, however, is the force exerted by the string on the bridge and the consequent emission of sound by the vibrating violin body. The spectrum transformation function relating the "internal" (bowing-point) spectrum to the "external" (room-average) spectrum must thus be considered in two parts. The first part relates the bowing-point velocity to the bridge force, while the second part converts this force spectrum into the one measured in the room.

The bowing-point velocity/bridge-force spectrum transformation function $T_{vF}(f)$ turns out, according to the simplest theory, to be $1/\sin(\pi f/f_\beta)$. The most striking consequence of this fact is that it exactly cancels out the notches in the simple formula for E
words, at those frequencies for which there is supposedly no velocity signal at all at the bowing point, the

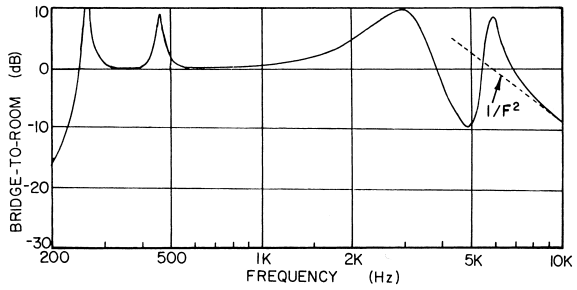


FIGURE 19 Simplified transformation function connecting the violin bridge drive-force spectrum envelope with the room-average spectrum envelope.

transformation function is so enormously effective that it apparently "creates" a drive force at the bridge! Curiously enough, although the stick-slip versions of both the velocity-spectrum and force-transformation function have been common knowledge for over a century, serious attempts to resolve the paradox have been made only recently. Much of the necessary information is currently available, but it has not been fitted together into a coherent whole. For present purposes, it will suffice to remark that the general trend of the force spectrum envelope of the bridge drive force is roughly constant at low frequencies, and it falls away fairly quickly for frequencies above a breakpoint that is determined in part by the distance of the bowing point from the bridge.

Figure 19 outlines the main behavior of the bridge-force-to-room transformation function. This must of course be evaluated for a drive force whose direction lies roughly parallel to the plane of the violin's top plate and tangent to the curved paths of the string anchorages on the bridge (see Fig. 17). To a first approximation the trend line is horizontal, in agreement with the general assertions made about the force-to-room transformation in connection with the plucked and struck string instruments. However, there is a very rapid weakening in the radiating ability of the body in the low-frequency region below a strong peak near 260 Hz. This radiative transformation peak and associated loss of low-frequency efficacy (whose cognate on the guitar falls at 85 Hz and below) is associated with a joint vibrational mode of the elastic-walled body cavity and the Helmholtz resonator formed by this air cavity and the apertures in it provided by the f holes (see Fig. 17 for details and terminology).

There is a second radiativity peak just below 500 Hz on a violin. This one is associated with the resonant response of a body mode in which the top plate vibrates (chiefly on the bass-bar side) in a sort of twisting motion having a quasi-fulcrum at the position of the sound post. The back plate is also in vigorous motion, being coupled to the top plate by the sound post.

(In the guitar, this radiativity peak is relatively unimportant. While it too has a body mode in which the bridge rocks strongly, the lack of a bass bar and sound post makes for a vibrational symmetry that gives a very small radiation of sound. Furthermore, the lowness of a guitar bridge means that this mode is only very weakly driven by a vibrating string.)

The violin has two more strong radiativity peaks. One is found near 3000 Hz, and the other near 6000 Hz, beyond which the radiativity falls as $1/f^2$ or faster. The first of these peaks is determined by a bridge-plus-body mode in which the predominant motion is a rocking of the top part of the bridge about its waist. The second peak belongs to a mode in which there is a sort of bouncing motion (normal to the plane of the top plate) of the upper part of the bridge on the bent "legs" connecting its waist to its feet. Analogs to these peaks do not exist on the guitar.

There are many additional resonance-related peaks and dips in the transformation function besides those described previously and indicated in Fig. 19. For a violin these are spaced (on the ge) only about 35 Hz apart, and they are proportionally closer on larger members of the family. We have dealt explicitly here only with those of major acoustical and musical importance whose positions along the frequency axis are well established for each family of bowed instruments. It is an important part of a fiddle maker's skill to place these selected peaks in their correct frequency positions. He must also properly proportion the interactions of the various parts of an instrument (e.g., by suitably dimensioning and placing the soundpost).

D. The Violin's Measured Room-Average Spectrum and Its Implications

Figure 20 shows the room-average spectra of all chromatic notes between a violin's bottom G_3 and the $A_4^{\#}$ that lies

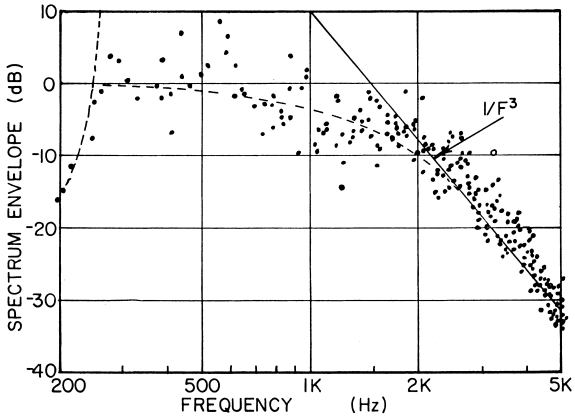


FIGURE 20 Measured room-average spectra of violin notes of the chromatic scale between G_3 and $A_4^{\#}$.

somewhat more than an octave above. One feature calls instant attention to itself in the spectrum envelope implied by these data: This measured envelope is similar to that of soprano wind instruments and also the piano in that the envelope is roughly uninform at low frequencies and falls away at the rate of about $1/f^3$ at high frequencies. The similarity is closest between the violin and the wind instruments, because the breakpoint between the low- and high-frequency regions lies in all cases near 1500 Hz!

Comparison of the envelope shape from Fig. 20 with the transformation function of Fig. 19 causes an initial feeling of surprise. Figure 19 shows a strong radiativity peak around 3000 Hz and another one near 450 Hz. Neither one of these shows up in the spectral envelope of the radiated sound. While no one seems to have worked out the details yet, the basic explanation has already been met among the wind instruments. Efficacy of radiation generally means a lowering of the resonance peaks that operate the primary excitation controller (reed or bow). As a result, there is less energy produced at the radiativity peaks than elsewhere, thus offsetting the increased emission of the enegy that is produced. However, the fact that all parts of the generated spectrum are strongly interconnected by heterodyne action makes it impossible to make detailed predictions of what will happen from general principles alone.

More detailed comparison of the radiativity curve and the spectrum envelope shows further evidence that their relationship is not simple: The strong radiativity peak at 260 Hz differs from the others in that it does appear to stregthen the radiated tonal components that coincide with it. Furthermore, the rapid decrease in radiativity below 250 Hz is reflected in a rapid loss of power in the corresponding components of the violin's tone. We also see (Fig. 20) hints of a strong emission of sound in the regions around 400 Hz and clear indications of even stronger emission around 550 Hz, despite the fact that there are no prominent resonances to be found at these frequencies in the violin's modal collection. Other hints of systematic fine structure in the observed spectrum are tantalizingly visible in the present data, hints that strengthen and weaken surprisingly when the data are displayed in different ways. As remarked earlier, much more remains to be done to elucidate the detailed origins of the violin's spectral envelope, as is the case with many other features of its acoustical behavior. Meanwhile, clues as to what sorts of phenomena are to be expected may be looked for in the spectral relations among the components of the brass instrument pedal tones.

The musical interpretability of the bowed string sound is made yet more difficult by the fact that the human auditory system is readily able to recognize the tonal influences of all the resonances displayed in Fig. 19, along with several other less well-marked or invariable ones belonging to this complex system. This is despite the fact that they are not visible in the measured spectrum. Physicists must always remember in cases like this that while the ear does not analyze sounds in the ways most readily chosen by laboratory scientists, it must in the final analysis act upon whatever pieces of physical data offer themselves, many of which can be at least listed for the scientist's serious consideration, even though they may be difficult for him to measure.

Claims are made from time to time that "the secret of Stradivari" has been discovered. Such claims arise in part because of a sometimes unrecognized conflict between the remarkably effective but subliminal routines of musical listening and the highly intellectualized activities of a laboratory researcher, and in part because of everyone's romantic desire to create a better instrument. Each discover proclaims some "truth" that he has found. If the "scientific" discoverer is often less guarded in his claims than in his craftsman or musician counterpart, it is because he often knows only one aspect of the primary oscillation problem or of the vibration/radiation aspects of the net sound production process. Moreover, he is not subjected to the discipline of successful practice in the real-world fields of instrument making or musical performance, where partial success is often equivalent to failure!

VI. THE APTNESS OF INSTRUMENTAL SOUNDS IN ROOMS

The diverse musical instruments that we have studied share a remarkable number of properties. Let us list some of these and attempt to relate them to the ways in which they provide useful data to the auditory processor.

All of the standard orchestral instruments generate sounds that are (note by note) made up of groups of sinusoids whose frequencies are whole-number multiples of some repetition rate. We might ask, at least for the plucked or struck string instruments, whether it is an accident that this should be so, since it comes about (and only approximately at that) via the choice of thin, elongated, uniform wires as the primary vibrating object. Why should such vibratiors take precedence over vibrating plates or membranes, or even over wires of nonuniform cross section? In the case of the wind and bowed instruments (including the singing voice), self-sustained oscillations are possible only under conditions where the resulting spectrum is of the strictly harmonic type. Here, then, the traditional instrument maker has no choice: It is impossible for him to provide inharmonic sound sources.

For the moment, the question remains partly open as to why the harmonic-type instruments are dominant. We

have however been given a strong hint by the observation that the auditory processor treats such aggregations of spectral components in a special way—it perceives each such grouping as an individual, compact tone. It can also distinguish several such tones at the same time, and even recognize well-marked relationships between them (such as the octave or the musical fifth).

The problem remains, however, as to whether the recognizability of individual harmonic groups can survive the room's transmission path. Regardless of the complexity of transmission of amplitude or phase, the frequencies of the components radiated from an instrument arrive unaltered at the listener's ear. It is an easily verified fact that the pitch of a harmonic complex is almost rigorously established by the harmonic pattern of its component frequencies (which determine in a mathematically unique way its repetition rate), rather than by the amplitudes of these components. In other words, as long as even a few of the partials of each instrument's tone detectably arrive at the listener's ear, the pitch (music's most important attribute) is well established.

Great emphasis has been laid throughout this article on the fact that each instrument is constructed in such a way that all of its notes share a common spectrum envelope. It has also been pointed out that for the keyboard instruments at least, it is a matter of considerable difficulty to achieve such an envelope. The structure of the brass instruments, on the other hand, almost guarantees a well-defined envelope; and even though it is possible to build woodwind instruments that lack an envelope, many things become easier if one is arranged for them. Finally, the guitarist and the violinist were found to have instruments that inherently tend to produce a spectral envelope, but one whose breakpoint and high-frequency slope can be influenced by the player via his choice of plucking or bowing point.

What are the perceptual reasons for these instruments to have evolved to produce a well-defined spectral envelope? This question can be answered at least in part by the facts of radiation acoustics and musical perception in rooms. It takes only a very limited collection of auditory samples of transmission-distorted data from an instrument for the listener to form an impression of the breakpoint and high-frequency slope, and so (even for a single note in a musical passage) to permit him to decide which of the instruments before him has produced it.

A question that is less easy to decipher is why the instruments seem to have very nearly the same envelopes. A partial explanation is to be found in the observation that the bulk of the available acoustical energy in a tone is allocated to the first four or five partials, which puts these energy packages into a set of independent critical bands, thus maximizing the net loudness. A less obvious explanation is that the high-frequency rolloff may be a way of preventing excessive tonal roughness of the sort that comes about when too many harmonics find themselves in the same critical band. For example, harmonics 7, 8, and 9 will contribute some roughness because they all lie within the 25% bandwidth of significant mutual interaction. This explanation is not adequate, however. It turns out that critical-band-induced roughnesses of this sort are not strongly active, whereas (for the soprano instruments at least) an insufficient rolloff rate tends to produce a quite unacceptable tone color.

While much remains to be done to fully clarify questions of the sort raised in the preceding paragraph, we can find hints as to where the answers may be sought. Psychoacousticians have shown the existence of a tonal attribute known as *sharpness* (which is what its German originator calls it in English, but edginess, or harshness, would be a better term). This attribute may be calculated dependably from the spectral envelope of a sound, its power level, and the frequency range in which its components are found. We may summarize the calculation method thus. The total loudness N perceived by the listener is given by the integral of a loudness density function $n(z)$ that is a perceptual cognate of the product of the physicist's spectral envelope $E(f)$ and the level of the acoustical signals received by the listener:

$$N = \int n(z)\,dz. \tag{18}$$

It also takes into account varying amounts of interaction between spectral components that are not widely separated in frequency. Here the variable z is the transformation of the ordinary frequency axis into a perceptual coordinate such that increments of one unit of z correspond to the width of one critical band. The sharpness, S, is then calculated as an overlap integral of $n(z)$ and a sharpness weighting function $g(x)$, given by

$$S = \frac{\text{const}}{\ln(N/20 + 1)} \int n(z)g(z)\,dz. \tag{19}$$

The sharpness weighting function $g(z)$ is small at low values of z, and it rises rapidly for values of z that correspond to frequencies above 1000 Hz. It should probably not be taken as accidental that the most important contributions to the sharpness integral arise above 1000 Hz, in the region where the primary receptors are beginning to randomly misfire relative to their mechanical stimuli.

Figure 21a shows the function $n(z)$ calculated for a harmonic tone having a fundamental frequency f_0 of 200 Hz, a spectral envelope with breakpoint frequency 1500 Hz,

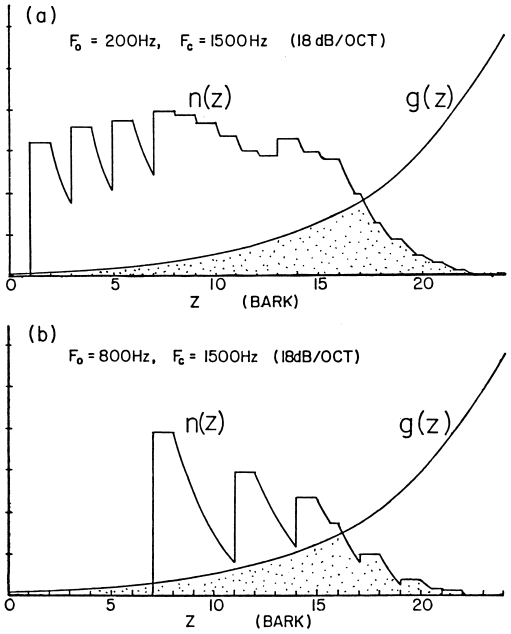


FIGURE 21 (a) Loudness density function $n(z)$ calculated for harmonic tones based on 200 Hz, and having a spectral envelope with 1500 Hz break frequency and $1/f^3$ high frequency rolloff. Also shown is the sharpness function $g(z)$. (b) Similar curves, for a tone having the same spectral envelope but with an 800 Hz fundamental frequency.

and $1/f^3$ high-frequency rolloff. Also plotted is the sharpness function $g(z)$. Qualitatively speaking, we may understand the net sharpness Σ as being related to the area of the shaded region lying between the z axis and the two curves. Figure 21b similarly illustrates the case of a tone belonging to the same spectral envelope and fundamental frequency 800 Hz.

We find by direct electronic synthesis (or by sounding a specially made laboratory wind instrument) that an unpleasant harshness is attributed to tones having a raised breakpoint or reduced rate of high-frequency falloff relative to the one we described previously. Furthermore, the tone is generally pronounced to lack piquancy and to be somewhat dull and muffled when the envelope has a lowered breakpoint or steepened falloff.

It is not difficult to expect from the general nature of Fig. 21 that instruments built to play in the alto, tenor, and bass ranges will be very little influenced by the sharpness phenomenon, freeing them (in accordance with observation and experiment) from the constraints that appear to hold for the soprano instruments. On the other hand, treble instruments (having breakpoint frequencies of 2000 Hz or above) are found to have a great deal of sharpness regardless of the high-frequency envelope slope. In the orchestra these instruments are rarely used, and then only for special purposes. Quantitative study of the relation of sharpness, loudness, and spectrum envelope for instruments in various pitch ranges is in its infancy, but already a considerable amount of consistency is apparent.

One flaw is to be noticed in the apparently coherent picture that has been sketched above: A most important musical instrument, the piano, seems not to provide the otherwise universal spectral envelope. Here the break frequency turns out to lie near 800 Hz rather than 1500 Hz. However, it is a simple matter to electronically refilter a set of recorded piano tones to move the breakpoint to 1500 Hz without changing anything else. Listening tests on such a modified (normalized) sound show at once that the tone does not so much become harsh as that the pounding of the hammers becomes obtrusive. Readers who have listened to the actual sound of the early-nineteenth-century pianoforte will have heard a mild form of the same kind of hammer heard a mild form of the same kind of hammer clang. (Do not count on a recording to inform you, because many recordings have been so tampered with that nothing can be learned from them.) Apparently, pianos have evolved away from an original design based on the harpsichord (where the continuous-spectrum impulsive hammer sound is not produced, and the spectrum envelope is essentially of the familiar type) to one in which one tonal virtue is sacrificed to avoid a serious flaw.

The perceptual symbiosis that exists between a musical instrument and the concert hall in which it is played can be illustrated further by considering the details of the primary radiation processes whose signals are compiled in making a room-average spectrum. For every instrument family, the spectrum envelope of the sounds radiated in some particular direction (in reflection-free surroundings) differs significantly from that radiated in some other direction. In many cases smoothly varying discrepancies between tow such envelopes can amount to as much as 40 dB. We also find that the signals from microphones placed at various positions close to an instrument have peculiar and highly irregular spectra that have no easily recognizable relationship with more thoughtfully obtained spectra. We have already seen what happens when many individual samples of the more distant version of the sound are combined into a room average: a reliable picture of it emerges. Our hearing mechanism and the surrounding hall join to accomplish just this task. The concert halls in which we normally listen to music offer many reflections, which means that data concerning all aspects (literally!) of the emitted sound are made available to our auditory processors. The chief reason (from the point of view of music) why the room-average spectrum is important is that the ear actually can assemble the equivalent information by means of early-reflection processing and/or multiple-sample averaging via use of two-ear, moving-listener, moving-source

data collected over the span of several seconds. Almost none of this multiplicity of data is available for processing in reflection-free surroundings, which provides a significant hint as to why serious performers and listeners alike tend to dislike open-air music: It subjects them to auditory deprivation.

Despite the noise-reduction and harmonic distortion-free techniques of digital recording and the use of compact disks, many modern attempts at musical recording are frequently quite unsatisfactory. Recording engineers sometimes misuse their technical resources in an attempt to remove the confusion from the recorded sound by the use of reflection-free studios, partitions between instruments, and the "mixing down" and "filter enhancement" of signals from numerous highly directional microphones (each placed very close to its own instrument). These actions (which are increasingly resented by performing classical musicians) produce distortion of the primary musical data when they do not eliminate them altogether. On the other hand, recordings of the sort made in the 1950s and 1960s using two or three microphones properly placed in a good concert hall have never been surpassed, at least in the informed judgement of those listeners to classical music whose experience has been gained largely by actual concert-going. In short, for music we need and enjoy all of the data from our instruments, instruments that have evolved over several centuries to communicate their voices effectively in the environment of a concert hall.

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • ACOUSTICS, LINEAR • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, GENERAL • ULTRASONICS AND ACOUSTICS

BIBLIOGRAPHY

Benade, A. H. (1976). "Fundamentals of Musical Acoustics," Oxford Univ. Press, London and New York.

Benade, A. H. (1985). From instrument to ear in a room: Direct, or via recording. *J. Audio Eng. Soc.* **33,** 218–233.

Benade, A. H., and Kouzoupis, S. N. (1988). The clarinet spectrum: Theory and experiment. *J. Acoust. Soc. Am.* **83,** 292–304.

Benade, A. H., and Larson, C. O. (1985). Requirements and techniques for measuring the musical spectrum of a clarinet. *J. Acoust. Soc. Am.* **78,** 1475–1497.

Benade, A. H., and Lutgen, S. J. (1988). The saxophone spectrum. *J. Acoust. Soc. Am.* **83,** 1900–1907.

Caussé, R., Kergomard, J., and Lurton, X. (1984). Input impedance of brass musical instruments—Comparison between experiment and numerical models. *J. Acoust. Soc. Am.* **75,** 241–254.

Cremer, L. (1984). "The Physics of Violins" (J. S. Allen, translator). MIT Press, Cambridge, Massachusetts.

De Poli, A. (1991). "Representations of Musical Signals," MIT Press, Cambridge, MA.

Griffith, N., and Todd, P. M. (1999). Musical Networks: Parallel Distributed Perception and Performance, MIT Press, Cambridge, MA.

Hall, D. E. (1986). Piano string excitation, I. *J. Acoust. Soc. Am.* **79,** 141–147.

Hall, D. E. (1987). Piano string excitation: The question of missing modes. *J. Acoust. Soc. Am.* **82,** 1913–1918.

Hall, D. E. (1988). Piano string excitation: Spectra for real hammers and strings. *J. Acoust. Soc. Am.* **83,** 1627–1638.

Hutchins, C. M. (1983). A history of violin research. *J. Acoust. Soc. Am.* **73,** 1421–1440.

Marshall, K. D. (1985). Modal analysis of a violin. *J. Acoust. Soc. Am.* **77,** 695–709.

McIntyre, M. E., Schumacher, R. T., and Woodhouse, J. (1983). On the oscillations of musical instruments. *J. Acoust. Soc. Am.* **74,** 1345–1375.

Pierce, J. R. (1992). "Science of Musical Sound, Rev. Ed." Holt, New York.

Rossing, T. D., and Fletcher, N. H. (1998). "The Physics of Musical Instruments, 2nd Ed," Springer-Verlag, New York.

Sadie, S. (ed.) (1980). "The New Grove Dictionary of Music and Musicians, Macmillan, London, England.

Weinreich, G., and Kergomard, J. (1996). "Mechanics of Musical Instruments," Springer-Verlag, New York.



Nonlinear Dynamics

F. C. Moon
Cornell University

I. Introduction
II. The Undamped Pendulum
III. Nonlinear Resonance
IV. Self-Excited Oscillations: Limit Cycles
V. Stability and Bifurcations
VI. Flows and Maps: Poincaré Sections
VII. One-Dimensional Maps, Bifurcations, and Chaos
VIII. Fractals and Chaotic Vibrations
IX. Fractal Dimension
X. Lyapunov Exponents and Chaotic Dynamics
XI. The Lorenz Equations: A Model for Convection Dynamics
XII. Spatiotemporal Dynamics: Solitons
XIII. Controlling Chaos
XIV. Conclusion

GLOSSARY

Bifurcation Denotes the change in the type of long-time dynamical motion when some parameter or set of parameters is varied (e.g., as when a rod under a compressive load buckles—one equilibrium state changes to two stable equilibrium states).

Chaotic motion Denotes a type of motion that is sensitive to changes in initial conditions. A motion for which trajectories starting from slightly different initial conditions diverge exponentially. A motion with positive Lyapunov exponent.

Controlling chaos The ability to use the parameter sensitivity of chaotic attractors to stabilize any unstable, periodic orbit in a strange attractor.

Duffing's equation Second-order differential equation with a cubic nonlinearity and harmonic forcing $\ddot{x} + c\dot{x} + bx + ax^3 = f_0 \cos \omega t$.

Feigenbaum number Property of a dynamical system related to the period-doubling sequence. The ratio of successive differences between period-doubling bifurcation parameters approaches the number 4.669.... This property and the Feigenbaum number have been discovered in many physical systems in the prechaotic regime.

Fractal dimension Fractal dimension is a quantitative property of a set of points in an n-dimensional space that measures the extent to which the points fill a subspace as the number of points becomes very large.

Hopf bifurcation Emergence of a limit cycle oscillation from an equilibrium state as some system parameter is varied.

Limit cycle In engineering literature, a periodic motion that arises from a self-excited or autonomous system as in aeroelastic flutter or electrical oscillations. In dynamical systems literature, it also includes forced periodic motions (see also Hopf bifurcation).

Linear operator Denotes a mathematical operation (e.g., differentiation, multiplication by a constant) in which the action on the sum of two functions is the sum of the action of the operation on each function, similar to the principle of superposition.

Lorenz equations Set of three first-order autonomous differential equations that exhibit chaotic solutions.

This set of equations is one of the principal paradigms for chaotic dynamics.

Lyapunov exponents Numbers that measure the exponential attraction or separation in time of two adjacent trajectories in phase space with different initial conditions. A positive Lyapanov exponent indicates a chaotic motion in a dynamical system with bounded trajectories. (Sometimes spelled Liapunov).

Nonlinearity Property of an input–output system or mathematical operation for which the output is not linearly proportional to the input. For example, $y = cx^n$ $(n \neq 1)$, or $y = x\,dx/dt$, or $y = c(dx/dt)^2$.

Period doubling Sequence of periodic vibrations in which the period doubles as some parameter in the problem is varied. In the classic model, these frequency halving bifurcations occur at smaller and smaller intervals of the control parameter. Beyond a critical accumulation parameter value, chaotic vibrations occur. This scenario to chaos has been observed in may physical systems but is not the only route to chaos (see Feigenbaum number).

Phase space In mechanics, an abstract mathematical space with coordinates that are generalized coordinates and generalized momenta. In dynamical systems, governed by a set of first-order evolution equations; the coordinates are the state variables or components of the state vector.

Poincaré section (map) Sequence of points in phase space generated by the penetration of a continuous evolution trajectory through a generalized surface or plane in the space. For a periodically forced second-order nonlinear oscillator, a Poincaré map can be obtained by stroboscopically observing the position and velocity at a particular phase of the forcing function.

Quasi-periodic Vibration motion consisting of two or more incommensurate frequencies.

Saddle point In the geometric theory of ordinary differential equations, an equilibrium point with real eigenvalues with at least one positive and one negative eigenvalue.

Solitons Nonlinear wave-like solutions that can occur in a chain of coupled nonlinear oscillators.

Strange attractor Attracting set in phase space on which chaotic orbits move; an attractor that is not an equilibrium point or a limit cycle, or a quasi-periodic attractor. An attractor in phase space with fractal dimension.

Van der Pol equation Second-order differential equation with linear restoring force and nonlinear damping, which exhibits a limit cycle behavior. The classic mathematical paradigm for self-excited oscillations.

DYNAMICS is the mathematical study of the way systems change in time. The models that measure this change include differential equations and difference equations, as well as symbol dynamics. The subject involves techniques for deriving mathematical models as well as the development of methods for finding solutions to the equations of motion. Such techniques involve both analytic methods, such as perturbation techniques, and numerical methods.

I. INTRODUCTION

In the classical physical sciences, such as mechanics or electromagnetics, the methods to derive mathematical models are classified as dynamics, advanced dynamics, Lagrangian mechanics, or Hamiltonian mechanics. In this review, we discuss neither techniques for deriving equations nor the specific solution methods. Instead, we describe some of the phenomena that characterize how nonlinear systems change in time, such as nonlinear resonance, limit cycles, coupled motions, and chaotic dynamics.

An important class of problems in this subject consists of those problems for which energy is conserved. Systems in which all the active forces can be derived from a force potential are sometimes called conservative. A branch of dynamics that deals with such systems is called Hamiltonian mechanics.

The qualifier nonlinear implies that the forces (or voltages, etc.) that produce change in physical problems are not linearly proportional to the variables that describe the state of the system, such as position and velocity in mechanical systems (or charges and currents in electrical systems). Mathematically, the term linear refers to the action of certain mathematical operators L, such as are used in multiplication by a constant, taking a derivative, or an indefinite integral. A linear operator is one that can be distributed among a sum of functions without interaction, that is,

$$L[af(z) + bg(t)] = aL[f(t)] + bL[g(t)].$$

Nonlinear operators, such as those that square or cube a function, do not obey this property. Dynamical systems that have nonlinear mathematical models behave very differently from ones that have linear models. In the following, we describe some of the unique features of nonlinear dynamical systems.

Another distinction is whether the motion is bounded or not. Thus, for a mass on an elastic spring, the restoring forces act to constrain the motion, whereas in the case of a rocket, the distance from some fixed reference can grow

without bound. In this review, we discuss only bounded problems typically involving vibrating phenomena.

Mathematical models in dynamical systems generally take one of three forms: differential equations (or flows), difference equation (called maps), and symbol dynamic equations. Although the physical laws from which the models are derived are often second-order differential equations, the theory of nonlinear dynamics is best studied by rewriting these equations in the form of first-order equations. For example, Newton's law of conservation of momentum for a unit mass with one degree of freedom is usually written as a second-order differential equation:

$$\ddot{x} = F(x, \dot{x}, t). \tag{1}$$

In nonlinear dynamics one often rewrites this in the form

$$\dot{x} = y, \quad \dot{y} = F(x, y, t). \tag{2}$$

The motion is then viewed in phase space with vector components (x, y) corresponding to position and velocity. (In advanced dynamics, phase space is sometimes defined in terms of generalized position coordinates and generalized momentum coordinates.) For more complex problems, one studies dynamical models with differential equations in an N-dimensional phase space with N components $\{x_1(t), x_2(t), \ldots, x_i(t), \ldots, x_n(t)\}$, where the equation of motion takes the form

$$\begin{aligned} \dot{x} &= \mathbf{F}(\mathbf{x}, t) \\ x_1 &= x_1 \qquad x_2 \equiv \dot{x} = y \end{aligned} \tag{3}$$

using Eq. (2).

Difference equations or maps are also used in nonlinear dynamics and are sometimes derived or related to continuous flows in phase space by observing the motion or state of the system at discrete times, that is, $x_n \equiv x(t_n)$. In distinction to Eq. (3), the subscript refers to different times or different events in the history of the system. First- and second-order maps have the following forms:

$$x_{n+1} = f(x_n) \tag{4a}$$

or

$$\begin{aligned} x_{n+1} &= f(x_n, y_n) \\ y_{n+1} &= g(x_n, y_n) \end{aligned} \tag{4b}$$

Examples are given later in this article.

Another model is obtained when the variable X_n is restricted to a finite set of values, say (0, 1, 2). In this case, there is no need to think in terms of numbers because one can make a correspondence between (0, 1, 2) and any set of symbols such as $(a_1, a_2, a_3) =$ (L, C, R) or (R, Y, B). Thus, in some systems we may be interested only in whether the particle is to the left (L), right (R), or in the center (C) with respect to some reference. We can also label states with colors, such as red (R), yellow (Y), or blue (B). The evolution of a system is then expressed in the form

$$a_{n+1} = h(a_n). \tag{5}$$

Here, however, $h(a_n)$ may not be an explicit algebraic expression but a rule that may incorporate inequalities. For example, suppose that $x(t_n)$ is the position of some particle at time t_n. Then one could have

$$\begin{aligned} a_{n+1} &= L \quad \text{if } x_n < 0 \\ a_{n+1} &= R \quad \text{if } x_n \geq 0. \end{aligned}$$

An equilibrium solution might be LLLL..., whereas a periodic motion has the form RRLR-RLRRL..., or LRLRLR....

For a given physical system, one can use all three types of models.

II. THE UNDAMPED PENDULUM

A. Free Vibrations

A classical paradigm in nonlinear dynamics is the circular motion of a mass under the force of gravity (Fig. 1). A balance equation between the gravitational torque and the rate of change of angular momentum yields the nonlinear ordinary differential equation

$$\ddot{\theta} + (g/L)\sin\theta = 0, \tag{6}$$

where g is the gravitational constant and L the length of the pendulum. A standard approach to understanding the dynamics of this system is to analyze the stability of motion of the linearized equations about equilibrium positions.

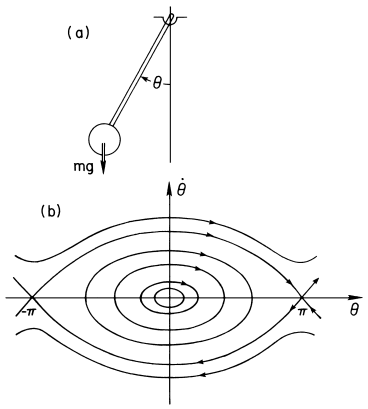


FIGURE 1 (a) The classical pendulum under the force of gravity. (b) Phase plane sketch of motions of the pendulum showing solutions near the origin (center) and solution near $\theta = \pm\pi$ (saddle point).

Using the form of Eq. (2) or (3) one has

$$\dot{\theta} = \Omega, \quad \dot{\Omega} = -\omega_0^2 \sin\theta \tag{7}$$

where

$$\omega_0^2 \equiv g/L.$$

Equilibrium points of Eq. (3) are defined by $\mathbf{F}(\mathbf{x}_e) = 0$. In the example of the pendulum, $\mathbf{x} = (\theta, \Omega)$ and $\theta_e = \pm m\pi$, $\Omega_e = 0$. Because the torque is periodic in θ, we can restrict θ to $-\pi < \theta \leq \pi$. In a linearized analysis, we define a perturbation variable $\varphi = \theta - \theta_e$ so that $\sin\theta$ is replaced by $\pm\varphi$, depending on whether $\theta_e = 0$ or π. About $\theta_e = 0$, one finds that the linearized motion is oscillatory (i.e., $\theta(t) = A\ \sin(\omega_0 t + B)$, where A and B are determined from initial conditions). The motion in the phase plane (θ, Ω) takes the form of an elliptic orbit with clockwise rotation (Fig. 1). Such motion is known as a center. The motion about $\theta_e = \pm\pi$ can be shown to be an unstable equilibrium point, known as a saddle, with trajectories that are also shown in Figure 1. (One should note that the saddles at $\theta_e = \pm\pi$ are physically the same.) Using the conservation of linearized system qualitatively represent those of the nonlinear system. These local qualitative pictures of the nonlinear phase plane motion can often be pieced together to form a global picture in Figure 1. The trajectory separating the inner orbits (libration) from the outer or rotary orbit is known as a separatrix. For small motions the period of oscillation is $2\pi/\omega_0$ or $2\pi(L/g)^{\frac{1}{2}}$. However, the period of libration increases with increasing amplitude and approaches infinity as the orbit approaches the separatrix. The dependence of the free oscillation period or frequency on the amplitude is characteristic of nonlinear systems.

III. NONLINEAR RESONANCE

A classical model for nonlinear effects in elastic mechanical systems is a mass on a spring with nonlinear stiffness. This model is represented by the differential equation (known as Duffing's equation)

$$\ddot{x} + 2\gamma\dot{x} + \alpha x + \beta x^3 = f(t). \tag{8}$$

This equation can also be used to describe certain nonlinear electrical circuits. When the linear damping term and external forcing are zero (i.e., $\gamma = f = 0$), the system is conservative and the nonlinear dynamics in the $(x, \dot{x} = y)$ phase plane can exhibit a number of different patterns of behavior, depending on the signs of α and β. When $\alpha, \beta > 0$. The system has a single equilibrium point, a center, where the frequency of oscillation increases with amplitude. For $\alpha > 0$ and $\beta < 0$, the frequency decreases

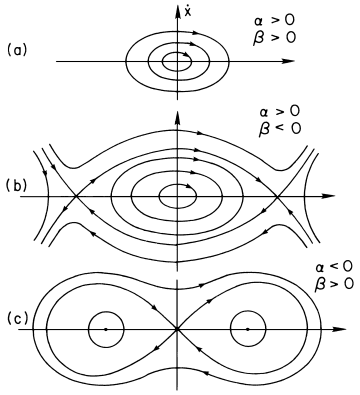


FIGURE 2 Phase plane motions for an oscillator with a nonlinear restoring force [Duffing's equation (8)]. (a) Hard spring problem, $\alpha, \beta > 0$. (b) Soft spring problem, $\alpha > 0$, $\beta < 0$. (c) Two-well potential problem, $\alpha < 0$, $\beta > 0$.

with amplitude (i.e., the period increases as in the pendulum) and the motion is unbounded outside the separatrix. For $\alpha < 0$ and $\beta > 0$, there are three equilibria: two stable and one unstable (a saddle), as in Figure 2c. Such motions represent the dynamics of a particle in a two-well potential.

Forced vibration of the damped system [Eq. (8)] represents an important class of problems in engineering. If the input force is oscillatory (i.e., $f = f_0 \cos\omega t$), the response of the system $x(t)$ can exhibit periodic, subharmonic, or chaotic functions of time. A periodic output has the same frequency as the input, whereas a subharmonic motion includes motions of multiple periods of the input frequency $2\pi/\omega$:

$$x(t) \sim A\ \cos[(n/m)\omega t + B]. \tag{9}$$

where n and m are integers. When the motion is periodic, the classic phenomenon of hysteretic nonlinear resonance occurs as in Figure 3. The output of the system has a different response for increasing versus decreasing forcing frequency in the vicinity of the linear natural frequency $\sqrt{\alpha}$. Also, the dotted curves in

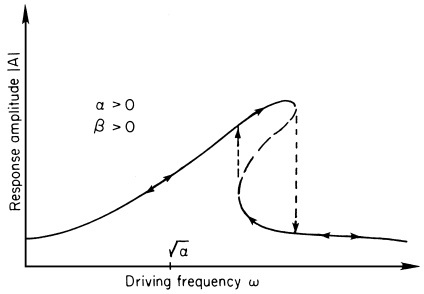


FIGURE 3 Nonlinear resonance for the hard spring problem: response amplitude versus driving frequency.

Figure 3 represent unstable motions that result in jumps in the response as frequency is increased or decreased. However, the output motion may not always be periodic, as Figure 3 implies, and may change to a subharmonic or chaotic motion depending on the parameters $(\gamma, \alpha, \beta, f_0, \omega)$. The multiplicity of possible solutions is not often pointed out in more classical treatments of nonlinear oscillations. Chaotic vibrations are discussed in the following.

IV. SELF-EXCITED OSCILLATIONS: LIMIT CYCLES

Dynamic systems with both sources and sinks for energy comprise an important class of nonlinear phenomena. These include mechanical systems with relative motion between parts, fluid flow around solid objects, biochemical and chemical reactions, and circuits with negative resistance (created by active electronic devices such as operational amplifiers or feedback circuits), as shown in Figure 4. The source of energy may create an unstable spiral equilibrium point while the source of dissipation may limit the oscillation motion to a steady motion or closed orbit in the phase space, as shown in Figure 5. The classical model for this limit cycle phenomena is the so-called Van der Pol equation given by

$$\ddot{x} - \gamma\dot{x}(1 - \beta x^2) + \omega_0^2 x = f(t). \tag{10}$$

When $f(t) = 0$, the system is called autonomous, and the origin is the only equilibrium point in the phase plane, that is, $(x, \dot{x} = \gamma) = (0, 0)$. This point can be shown to

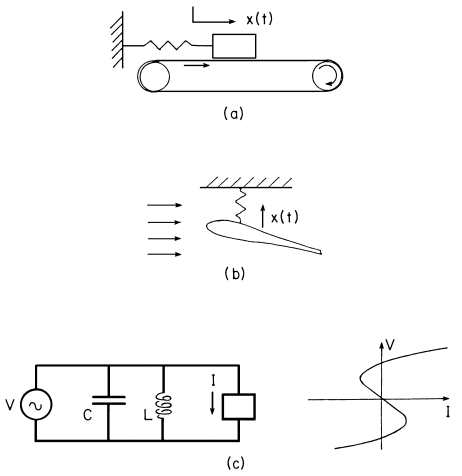


FIGURE 4 Sources of self-excited oscillations. (a) Dry friction between a mass and a moving belt. (b) Aeroelastic forces on a vibrating airfoil. (c) Negative resistance in an active circuit element.

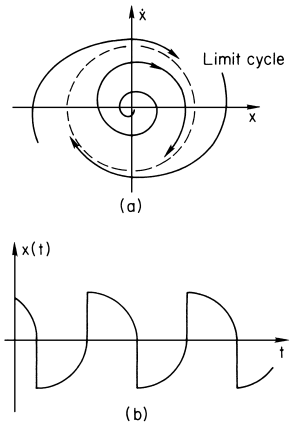


FIGURE 5 Phase plane portrait for a limit cycle oscillation. (a) Small γ [Eq. (10)]. (b) Relaxation oscillations, large γ [Eq. (10)].

be an unstable spiral when $\gamma > 0$. When γ is small, the limiting orbit in a set of normalized coordinates ($\beta = \omega_0^2 = 1$) is a circle of radius 2. As shown in Figure 5a, solutions inside the circle spiral out and onto the limit cycle while those outside spiral inward and onto the limit orbit. The frequency of the resulting periodic motion for $\beta = \omega = 1$ is one radian per nondimensional time unit.

When γ is larger (e.g., $\gamma \sim 10$), the motion takes a special form known as a relaxation oscillation, as shown in Figure 5b. It is periodic but is not sinusoidal, that is, it includes higher harmonics. The system exhibits sudden periodic shifts in motion.

If periodic forcing is added to a self-excited system such as Eq. (10) (i.e., $f(t) = f_0 \cos \omega_1 t$), then more complicated motions can occur. Note that when a nonlinear system is forced, superposition of free and forced motion is not valid. Two important phenomena in forced, self-excited systems are mentioned here: entrained oscillation and combination or quasi-periodic oscillations. When the driving frequency is close to the limit cycle frequency, the output $x(t)$ may become entrained at the driving frequency. For larger differences between driving and limit cycle frequencies, the output may be a combination of the two frequencies in the form

$$x = A_1 \cos \omega_0 t + A_2 \cos \omega_1 t. \tag{11}$$

When ω_0 and ω_1 are incommensurate (i.e., ω_0/ω_1 is an irrational number), the motion is said to be quasi-periodic, or almost periodic. Phase plane orbits of Eq. (11) are not closed when ω_0 and ω_1 are incommensurate.

V. STABILITY AND BIFURCATIONS

The existence of equilibria or steady periodic solutions is not sufficient to determine if a system will actually behave

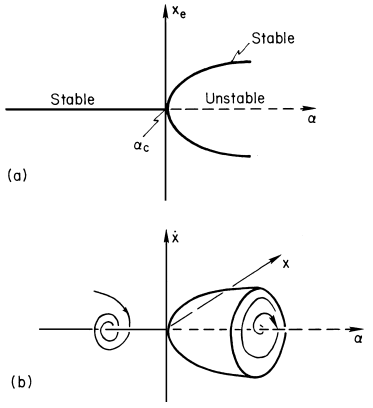


FIGURE 6 Bifurcation diagrams. (a) Pitchfork bifurcation, the transition from one to two stable equilibrium positions. (b) Hopf bifurcation, the transition from stable spiral to limit cycle oscillation.

that way. The stability of these solutions must also be checked. As parameters are changed, a stable motion can become unstable and new solutions may appear. The study of the changes in the dynamic behavior of systems as parameters are varied is the subject of bifurcation theory. Values of the parameters at which the qualitative or topological nature of the motion changes are known as critical or bifurcation values.

An example of a simple bifurcation is the equation for motion in a two-well potential [Eq. (8)]. Suppose we view α as a control parameter. Then in Eq. (8), the topology of the phase space flow depends critically on whether $\alpha < 0$ or $\alpha > 0$, as shown in Figure 2a and c, for zero damping and forcing. Thus $\alpha = 0$ is known as the critical or bifurcation value. A standard bifurcation diagram plots the values of the equilibrium solution as a function of α (Fig. 6a) and is known as a pitchfork bifurcation. When damping is present, the diagram is still valid. In this case, one stable spiral is transformed into two stable spirals and a saddle as α decreases from positive to negative values.

A bifurcation for the emergence of a limit cycle in a physical system is shown in Figure 6b. This is sometimes known as a Hopf bifurcation. Here, the equilibrium point changes from a stable spiral or focus to an unstable spiral that limits onto a periodic orbit.

VI. FLOWS AND MAPS: POINCARÉ SECTIONS

An old technique for analyzing solutions to differential equations, developed by Poincaré around the turn of the 20th century, has now assumed greater importance in the modern study of dynamical systems. The Poincaré section is a method to transform a continuous dynamical process in time into a set of difference equations of the form of Eq. (4b), known in modern parlance as a map. The study of maps obtained from Poincaré sections of flows is based on the theory that certain topological features of the motion in time are preserved in the discrete time dynamics of maps.

To illustrate how a Poincaré section is obtained, imagine that a system of three first-order differential equations of the form of Eq. (3) has solutions that can be represented by continuous trajectories in the Cartesian space (x, y, z), where $x_1(t) = x$, $x_2(t) = y$, and $x_3(t) = z$ (Fig. 7). If the solutions are bounded, then the solution curve is contained within some finite volume in this space. We then choose some surface through which the orbits of the motion pierce. If a coordinate system is set up on this two-dimensional surface with coordinates (ξ, η), then the position of the $(n + 1)$th orbit penetration (ξ_{n+1}, η_{n+1}) is a function of the nth orbit penetration through the solution of the original set of differential equations.

A period-one orbit means that

$$\xi_{n+1} = \xi_n, \ \eta_{n+1} = \eta_n.$$

A period-m orbit is defined such that

$$\xi_{n+m} = \xi_n, \eta_{n+m} = \eta_n.$$

Such orbits in the map correspond to periodic and subharmonic motions in the original continuous motion. On the other hand, if the sequence of points in the map seem to lie on a closed curve in the Poincaré surface, the motion is termed quasi-periodic and corresponds to the sum of two time-periodic functions of different incommensurate frequencies, as in Eq. (9).

Motions whose Poincaré maps have either a finite set of points (periodic or subharmonic motion) or a closed curve of points are known as classical attractors. A motion with a set of Poincaré points that is not a classical attractor and that has certain fractal properties is known as a strange attractor. Strange attractor motions are related to chaotic motions and are defined as follows.

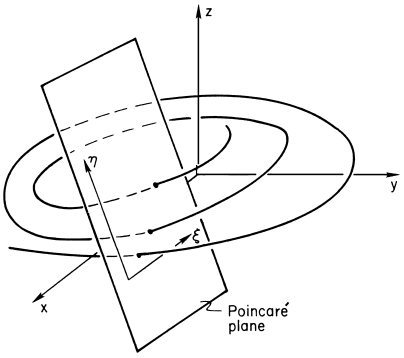


FIGURE 7 Poincaré section. Construction of a difference equation model (map) from a continuous dynamic model.



FIGURE 8 Experimental Poincaré map for chaotic motions of a particle in a two-well potential with periodic forcing and damping [Eq. (10)].

In certain periodically forced problems, there is a natural way to obtain a Poincaré section or map. Consider the damped mass with a nonlinear spring and time periodic force

$$\dot{x} = y$$
$$\dot{y} = -\gamma y - F(x) + f_0 \cos \omega t. \tag{12}$$

A Poincaré section can be obtained in this system by defining a third variable $z = \omega t$, where $0 \leq z < 2\pi$, so that the system is converted to a autonomous system of equations using $\dot{z} = \omega$. We also connect the planes defined by $z = 0$ and $z = 2\pi$ so that the motion takes place in a toroidal volume (Fig. 7). The Poincaré map is obtained by observing (x, y) at a particular phase of the forcing function. This represents a stroboscopic picture of the motion. Experimentally, one can perform the phase plane trace at a particular phase $z = z_0$ on a storage oscilloscope (Fig. 8).

VII. ONE-DIMENSIONAL MAPS, BIFURCATIONS, AND CHAOS

A simple linear difference equation has the form

$$x_{n+1} = \lambda x_n. \tag{13}$$

This equation can be solved explicitly to obtain $x_n = A\lambda^n$, as the reader can check. The solution is stable (i.e., $|x_n| \to 0$ as $n \to \infty$) if $|\lambda| < 1$ and unstable if $|\lambda| > 1$. The linear equation [Eq. (13)] is often used as a model for population growth in chemistry and biology. A more realistic model, which accounts for a limitation of resources in a given species population, is the so-called logistic equation

$$x_{n+1} = \lambda x_n (1 - x_n). \tag{14}$$

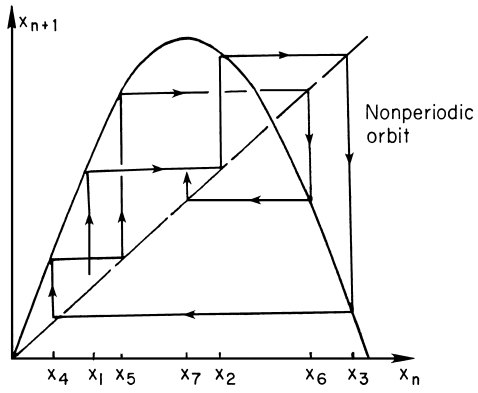


FIGURE 9 Graphical solution to a first-order difference equation. The example shown is the parabolic or logistic map.

This is a nonlinear difference equation that has equilibrium points $x = 0, 1$. One can examine the stability of nonlinear maps in the same way as for flows by linearizing the right-hand side of Eq. (14) about the equilibrium or fixed points.

The orbits of a solution to one-dimensional maps can be solved graphically by reference to Figure 9, in which the $(n + 1)$th value is reflected about the identity orbit (straight line). An orbit consists of a sequence of points $\{x_n\}$ that can exhibit transient, periodic, or chaotic behavior, as shown in Figure 10. These properties of solutions can be represented by the bifurcation diagram in Figure 11, where λ is a control parameter. As λ is varied, periodic solutions change character to subharmonic orbits of twice the period of the

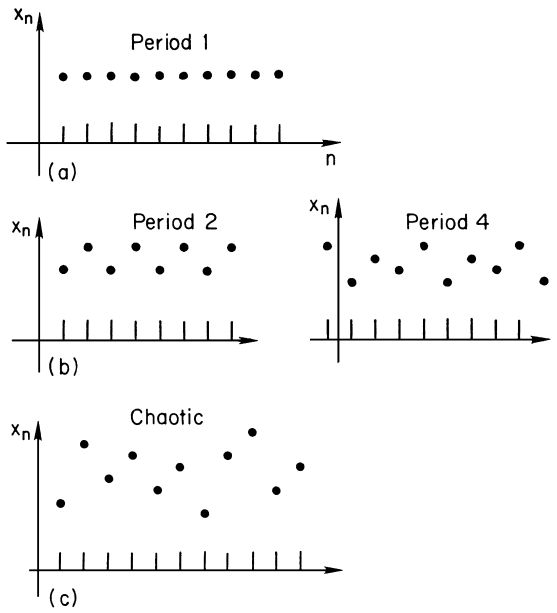


FIGURE 10 Possible solutions to the quadratic or parabolic map [Eq. (14)]. (a) Steady or period-one motion. (b) Period-two and period-four motions. (c) Chaotic motions.

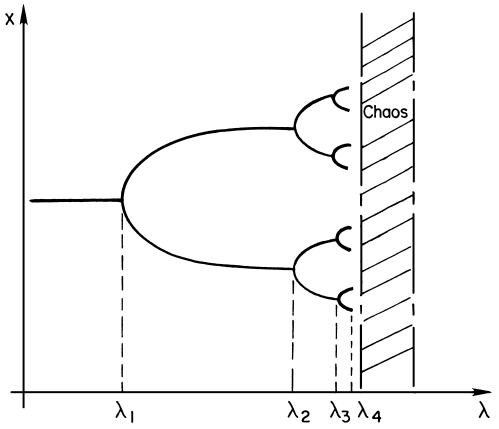


FIGURE 11 Period-doubling bifurcation diagram for a first-order nonlinear difference equation [Eq. (14)].

previous orbit. The bifurcation values of λ, $\{\lambda_n\}$ accumulate at a critical value at which non-periodic orbits appear. The sequences of values of λ at which period-doubling occurs has been shown by Feigenbaum to satisfy the relation

$$\lim[(\lambda_n - \lambda_{n-1})/(\lambda_{n+1} - \lambda_n)] \to 4.6692\ldots. \qquad (15)$$

These results have assumed great importance in the study of dynamic models in classical physics for two reasons. First, in many experiments, Poincaré sections of the dynamics often (but not always) reveal the qualities of a one-dimensional map. Second, Feigenbaum and others have shown that this period-doubling phenomenon is not only a prelude to chaos but is also universal when the one- dimensional map $x \to f(x)$ has at least one maximum or hump. Universal means that no matter what physical variable is controlled, it shows the same scaling properties as Eq. (15). This has been confirmed by many experiments in physics in solid- and fluid-state problems. However, when the underlying dynamics reveals a two-dimensional map, then the period-doubling route to chaos may not be unique.

A two-dimensional map that can be calculated directly from the principles of the dynamics of a ball bouncing on a vibrating platform under the force of gravity is shown in Figure 12a and b (Guckenheimer and Holmes have given a derivation). The difference equations are given by

$$x_{n+1} = (1 - \varepsilon)x_n + \kappa \sin y_n \qquad (16)$$

and

$$y_{n+1} = y_n + x_{n+1},$$

where x_n is the velocity before impact, y_n the time of impact normalized by the frequency of the vibrating table (i.e., $y = \omega t$, modulo 2π), and κ proportional to the amplitude of the vibrating table in Figure 12a. The parameter ε is proportional to the energy lost at each impact with the table. When the system is conservative, $\varepsilon = 0$, the Eqs. (16) are essentially a Poincaré map of the continuous motion obtained by observing the time of phase and velocity of impact when the ball hits the table. The first equation is a momentum balance relation before and after impact, whereas the second equation is found by integrating the free flight motion of the ball between impacts.

These equations have also been used to model an electron in an electromagnetic field. This map is sometimes known as the standard map.

In this problem, one can compare the difference between chaos in a conservative system ($\varepsilon = 0$) and chaos in a system for which there is dissipation. When $\varepsilon = 0$ (Fig. 12c), the map shows there are periodic orbits (fixed points in the map) and quasi-periodic motions, as evidenced by the closed map orbits. Islands of chaos exist for initial conditions starting near the saddle points of the

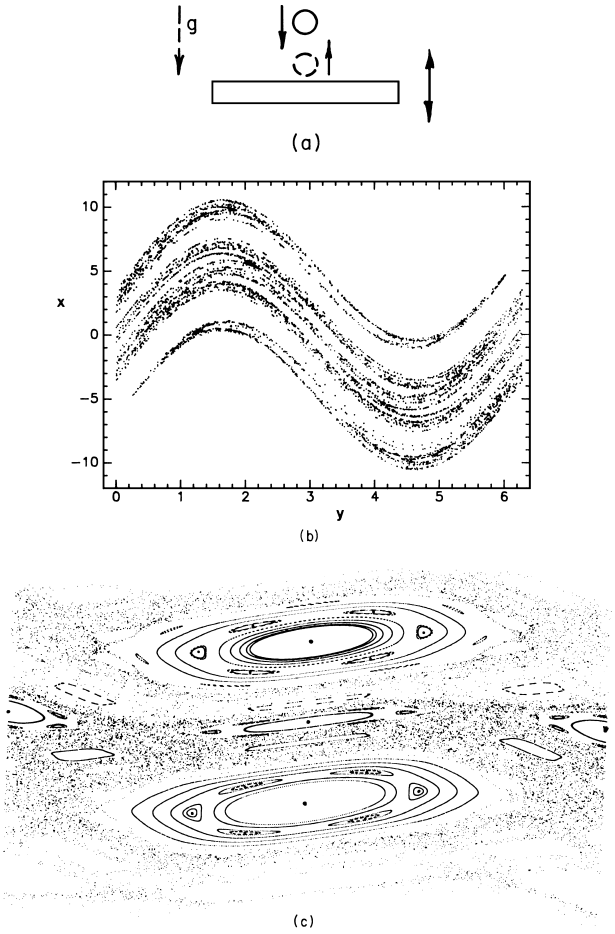


FIGURE 12 Dynamics of the second-order standard maps [Eq. (16)]. (a) Physical model of a ball bouncing in a vibrating table. (b) Iterations of the map with dissipation $\varepsilon = 0.4$, $\kappa \sim 6$ [Eq. (16)], showing fractal structure characteristic of strange attractors. (c) Iteration of the map for many different initial conditions showing regular and stochastic motions (no dissipation $\varepsilon = 0$, $\kappa \sim 1$).

map. In the dissipative case (Fig. 12b), the chaotic orbit shows a characteristic fractal structure but requires a much larger force amplitude κ. However, the forcing amplitude κ needed to obtain chaotic motion in the dissipative case is much larger than that required for chaos in the conservative case $\varepsilon = 0$.

VIII. FRACTALS AND CHAOTIC VIBRATIONS

One of the remarkable discoveries in nonlinear dynamics in recent years is the existence of randomlike solutions to deterministic differential equations and maps. Stochastic motions in nondissipative or conservative systems were known around the time of Poincaré. However, the discovery of such motions in problems with damping or dissipation was a surprise to many theorists and has led to experimental observations of chaotic phenomena in many areas of classical physics. Technically, a chaotic motion is one in which the solution is extremely sensitive to initial conditions, so much so that trajectories in phase space starting from neighboring initial conditions diverge exponentially from one another on the average. In a flow, this divergence of trajectories can take place only in a three-dimensional phase space. In a map, however, one can have chaotic behavior in a first-order nonlinear difference equation, as described in the logistic map example of Eq. (14).

In the dissipative standard map for the bouncing ball [Eq. (16)], chaotic solutions exist when impact energy is lost ($\varepsilon > 0$) for $\kappa \sim 6$. A typical long iterative map of such a solution is shown in Figure 12b. The iterates appear to occur randomly along the sets of parallel curves. If this solution is looked at with a finer grid, this parallel structure continues to appear. The occurrence of self-similar structure at finer and finer scales in this set of points is called fractal. Fractal structure in the Poincaré map is typical of chaotic attractors.

IX. FRACTAL DIMENSION

A quantitative measure of the fractal property of strange attractors is the fractal dimension. This quantity is a measure of the degree to which a set of points covers some integer n-dimensional subspace of phase space. There are many definitions of this measure. An elementary definition is called the capacity dimension. One considers a large number of points in an n-dimensional space and tries to cover these points with a set of N hypercubes of size ε. If the points were uniformly distributed along a linear curve, the number of points required to cover the set would vary as $N \sim 1/\varepsilon$ (Fig. 13). If the points were distributed on a two-

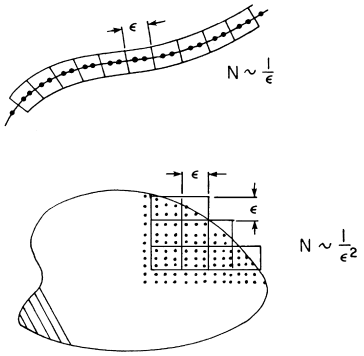


FIGURE 13 Definition of fractal dimension of a set of points in terms of the number of covering cubes $N(\varepsilon)$.

dimensional surface, then $N \sim \varepsilon^{-2}$. When the points are not uniformly distributed, one might find that $N \sim \varepsilon^{-d}$. If this behavior continues as $\varepsilon \to 0$ and the number of points increases, then we define the capacity dimension as

$$d = \lim_{\varepsilon \to 0} [\log N(\varepsilon) / \log(1/\varepsilon)]. \tag{17}$$

Other definitions exist that attempt to measure fractal properties, such as the information dimension and the correlation dimension. In the latter, one chooses a sphere or hypersphere of size ε and counts the number of points of the set in the sphere. When this is repeated for every point in the set, the sum is called the correlation function $C(\varepsilon)$. If the set is fractal, then $C \sim \varepsilon^d c$ or $d_c = \lim(\log C / \log \varepsilon)$, as $\varepsilon \to 0$. It has been found that $d_c \leq d$, where d is the capacity dimension.

As an example, consider the chaotic dynamics of a particle in a two-well potential. The equation of motion is given by

$$\ddot{x} + \gamma \dot{x} - \frac{1}{2} x(1 - x^2) = f_0 \cos \omega t. \tag{18}$$

This is a version of the Duffing equation [Eq. (8)]. It has been shown by Holmes of Cornell University that for chaos to exist, the amplitude of the forcing function must be greater a critical value, that is,

$$f_0 > [\gamma \sqrt{2} \cosh(\pi \omega / \sqrt{2})] / 3\pi\omega. \tag{19}$$

The regions of chaos in the parameter plane (f_0, ω) are shown in Figure 14, as determined by numerical experiments. The above criterion gives a good lower bound. Equation (18) is a model for the vibrations of a buckled beam. Experiments with chaotic vibrations of a buckled beam show good agreement with this criterion [Eq. (19)].

The fractal dimension of the Poincaré map of the chaotic motions of Eq. (18) depends on the damping γ. When γ is large ($\gamma \sim 0.5$), the dimension $d_c \sim 1.1$, and when γ is small ($\gamma \sim 0.01$), $d_c \sim 1.9$. The fractal dimension of the

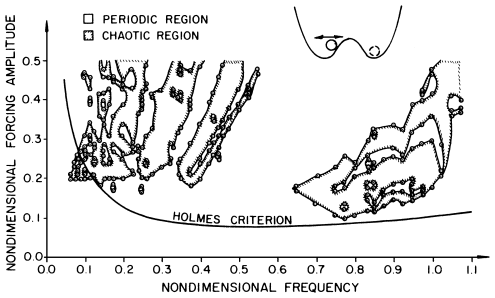


FIGURE 14 Regions of chaotic and regular motion for the two-well potential problem [Eq. (18)].

set of points in Figure 15a is close to $d_c = 1.5$. This means that the points do not cover the two-dimensional plane. This is evidenced by the voidlike structure of this chaotic map as it is viewed at finer and finer scales.

X. LYAPUNOV EXPONENTS AND CHAOTIC DYNAMICS

A measure of the sensitivity of dynamical motion to changes in initial conditions is the Lyapunov exponent. Thus, if two trajectories start close to one another, the distance between the two orbits $\Delta(t)$ increases exponentially in time for small times, that is,

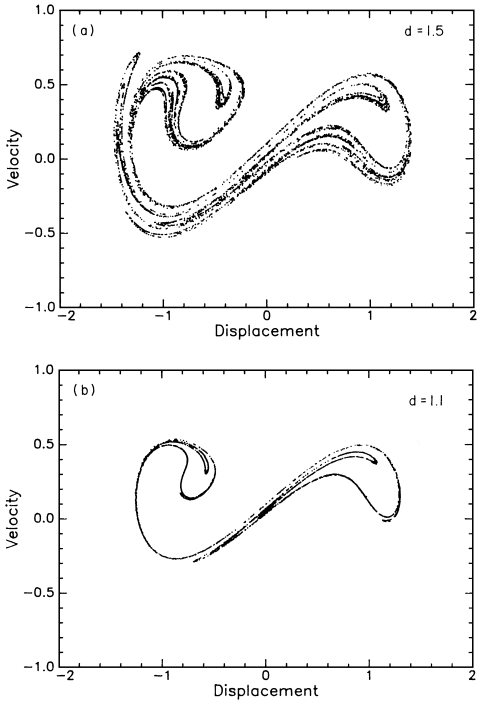


FIGURE 15 Poincaré maps of chaotic motions of the two well potential problem with fractal dimensions of (a) $d = 1.5$ and (b) $d = 1.1$ for two different damping ratios.

$$\Delta(t) = \Delta_0 2^{\lambda t}. \tag{20}$$

When λ is averaged over many points along the chaotic trajectory, λ is called the Lyapunov exponent. In a chaotic motion $\lambda > 0$, whereas for regular motion $\lambda \leq 0$.

In a two-dimensional map, one imagines a small circle of initial conditions about some point on the attractor (Fig. 16). If the radius of this circle is ε, then after several iterations of the map (say n), the circle may be mapped into an ellipse with principal axis of dimension $(2\varepsilon\mu_1^n,\ 2\varepsilon\mu_2^n)$; $\mu_1,\ \mu_2$ are called Lyapunov numbers, and the exponents are found from $\lambda_i = \log \mu_i$. If the system is dissipative, the area decreases after each iteration of the map (i.e., $\lambda_1 + \lambda_2 < 0$). If $\lambda_1 > \lambda_2$, then in a chaotic motion $\lambda_1 > 0$ and $\lambda_2 < 0$.

Thus, regions of phase space are stretched in one direction and contracted in another direction (Fig. 16). Iterating this stretching and contraction through the map eventually produces the fractal structure seen in Figure 15. Of importance to the understanding of such motions are concepts such as horseshoe maps and Cantor sets. Space does not permit a discussion of these ideas, but they may be found in several of the modern references on the subject.

The Lyapunov exponents $(\lambda_1,\ \lambda_2)$ can be used to calculate another measure of fractal dimension called the Lyapunov dimension. For points in a plane, such as a two-dimensional Poincaré map, this measure is given by

$$d_L = 1 - (\lambda_1/\lambda_2) = 1 + [\log \mu_1 / \log(1/\mu_2)]. \tag{21}$$

where $\lambda_1 > 0,\ \lambda_2 < 0$. This relation can be extended to higher dimensional maps.

XI. THE LORENZ EQUATIONS: A MODEL FOR CONVECTION DYNAMICS

As a final illustration of the new concepts in nonlinear dynamics, we consider a set of three equations proposed by Lorenz of MIT in 1963 as a crude model for thermal gradient induced fluid convection under the force of gravity. Such motions occur in oceans, atmosphere, home heating, and many engineering devices. In this model the variable x represents the amplitude of the fluid velocity stream function and y and z measure the time history of the temperature distribution in the fluid (a derivation has been

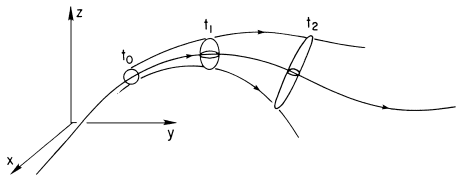


FIGURE 16 Sketch of sensitivity of motion to initial conditions as measured by Lyapunov exponents.

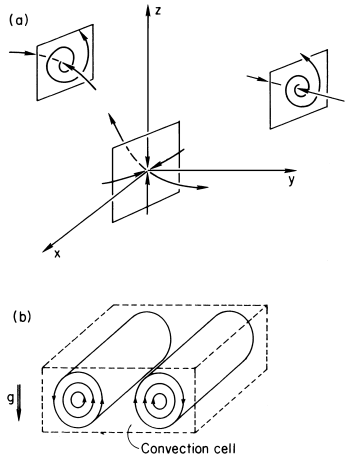


FIGURE 17 Sketch of local motion near the three equilibria for the Lorenz equations [Eq. (22)]. (a) As a model for thermo-fluid convection (b).

given by Lichtenberg and Lieberman). In nondimensional form, these equations become

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= xy - bz \end{aligned} \tag{22}$$

These equations would be linear were it not for the two terms xz and xy in the second and third equations. For those familiar with fluid mechanics, σ is a Prandtl number and r is similar to a Rayleigh number. The parameter b is a geometric factor. If $(\dot{x},\ \dot{y},\ \dot{z}) \equiv \mathbf{v}$ were to represent a velocity vector in phase space, then the divergence $\nabla \cdot \mathbf{v} = -(\sigma + b + 1) < 0$. This implies that a volume of initial conditions decreases as the motion moves in time. For $\sigma = 10,\ b = \frac{8}{3}$ (a favorite set of parameters for experts in the subject), there are three equilibia for $r > 1$ with the origin an unstable saddle (Fig. 17a). When $r > \sim 25$, the other two equilibria become unstable spirals (Fig. 17a) and a complex chaotic trajectory moves between regions near all three equilibria as shown in Figure 18.

An appropriate Poincaré map of this flow shows it to be nearly a one-dimensional map with period-doubling behavior as r is increased close to the critical value of $r = \sim 25$. The fractal dimension of this attractor has been found to be 2.06, which indicates that the motion lies close to a two-dimensional surface.

XII. SPATIOTEMPORAL DYNAMICS: SOLITONS

Nonlinear dynamics models can be used to study spatially extended systems such as acoustic waves, electrical transmission problems, plasma waves, and so forth. These

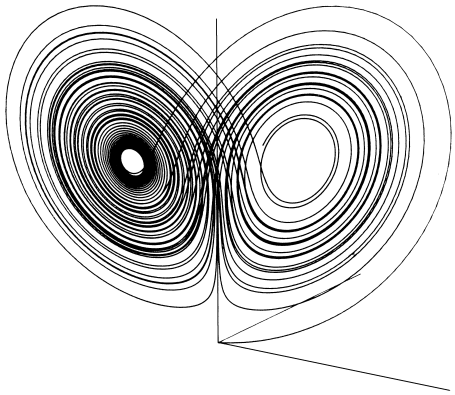


FIGURE 18 Trajectories of chaotic solution to the Lorenz equations for thermo-fluid convection.

problems have been modeled by using a linear chain of discrete oscillators with nearest neighbor coupling as shown in Figure 19. Of course, the limit of such models is continuum physics, such as acoustics or fluid mechanics, for which one uses partial differential equations in space and time. When the coupling between the oscillators is nonlinear, then a phenomena known as *soliton wave* dynamics can occur. Solitons are pulse-like waves that can propagate along the linear chain. Left- and right-moving waves can intersect and emerge as left and right waves without distortion (see Fig. 20). One example is the so-called Toda-lattice, in which the inter particle force is assumed to vary exponentially. (See Toda, 1989.)

$$\begin{aligned} &\ddot{x}_j + F(x_j - x_{j-1}) + F(x_j - x_{j+1}) = 0 \\ &F(x) = \beta[(\exp(bx) - 1)]. \end{aligned}$$

A classic problem of dynamics on a finite particle chain was posed by the famous physicist Eurico Fermi and two colleagues at Los Alamos in the 1950s. They had expected that if energy were placed in one spatial mode, then the nonlinear coupling would disperse the waves into the N classic vibration modes. However to their surprise, most of the energy stayed in the first spatial mode and eventually, after a finite time, all the energy returned to the original initial energy mode. This phenomenon is known as the *recurrence problem* in finite degree of freedom systems and is known as the Fermi–Pasta–Ulam problem. (See Gapanov-Grekhov and Rubinovich, 1992.)

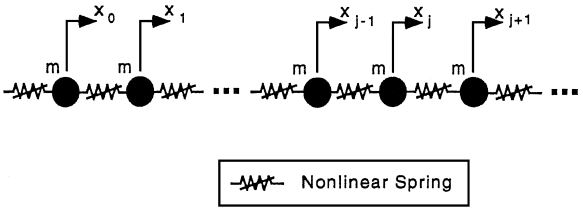


FIGURE 19 Chain of coupled nonlinear oscillators.

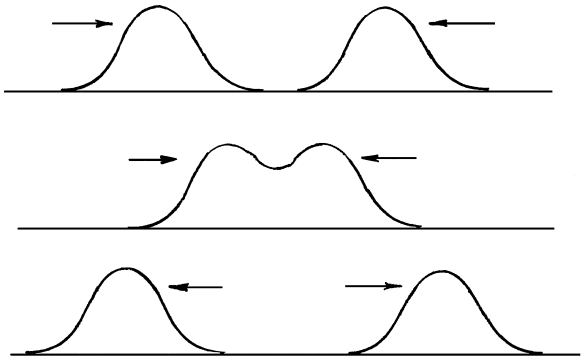


FIGURE 20 Soliton wave dynamics along a chain of nonlinear oscillator.

However, it is also now recognized that a different set of initial conditions could result in spatiotemporal stochasticity or *spatiotemporal chaos*. (See e.g., Moon, 1992.)

XIII. CONTROLLING CHAOS

One of the most inventive ideas to come out of modern nonlinear dynamics is the control of chaos. It is based on the concept that a system with a chaotic attractor may be used as a source of controlled periodic motions. This idea originated in the work of three University of Maryland researchers in 1990; E. Ott, C. Gregogi, and J. Yorke (OGY). (See Kapitaniak, 1996 for a review of this subject.) This example of *nonlinear thinking* has resulted in the design of systems to control chaotic modulation of losers, systems to control heart arythymias, and circuits to encrypt and decode information.

The idea is based on several premises.

1. The nonlinear system has a chaotic or strange attractor.
2. That the strange attractor is robust in some variation of a control parameter.
3. That there exists an infinite number of unstable periodic orbits in the strange attractor.
4. There exists a control law that will locally stabilize the unstable motion in the vicinity of the saddle points of the orbit map.

There are many variations of this OGY method, some based on the analysis of the underlying nonlinear map and some based on experimental techniques. An example of controlled dynamics is shown in Figure 21 from the author's laboratory. The vertical scale shows the Poincaré

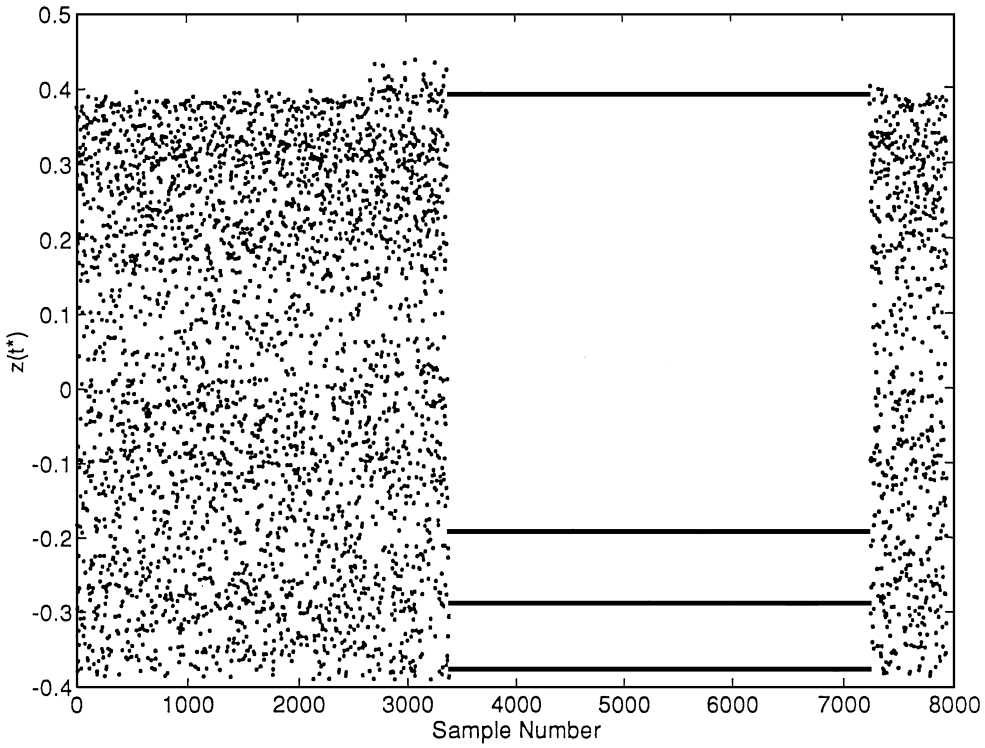


FIGURE 21 Poincaré map amplitude versus time for chaotic and controlled chaos of a period-four orbit for a two-well potential nonlinear oscillator.

map sampled output of a vibrating nonlinear elastic beam. The horizontal scale shows the time. The figure shows first a chaotic signal, then when control in initiated, a period four orbit appears, then chaos returns when the control is switched off. The control consists of a pulsed magnetic force on the beam. The control force is only active for a fraction of the period. The method uses the inherent parameter sensitivity of the underlying chaotic attractor to achieve control.

XIV. CONCLUSION

Dynamics is the oldest branch of physics. Yet, 300 years after the publication of Newton's Principia (1687), new discoveries are still emerging. The ideas of Newton, Euler, Lagrange, Hamilton, and Poincaré, once conceived in the context of orbital mechanics of the planets, have now transcended all areas of physics and even biology. Just as the new science of dynamics in the seventeenth century gave birth to a new mathematics (namely, the calculus), so have the recent discoveries in chaotic dynamics ushered in modern concepts in geometry and topology (such as fractals), which the 21st century novitiate in dynamics must master to grasp the subject fully. The bibliography lists only a small sample of the literature in dynamics. However, the author hopes that they will give the interested reader some place to start the exciting journey into the field of nonlinear dynamics.

SEE ALSO THE FOLLOWING ARTICLES

CHAOS • DYNAMICS OF ELEMENTARY CHEMICAL REACTIONS • FLUID DYNAMICS • FRACTALS • MATHEMATICAL MODELING • MECHANICS, CLASSICAL • NONLINEAR PROGRAMMING • VIBRATION, MECHANICAL

BIBLIOGRAPHY

Abraham, R. H., and Shaw, C. D. (1983). "Dynamics: The Geometry of Behavior," Parts 1–3. Aerial Press, Santa Cruz, CA.

Gaponov-Grekhov, A. V., and Rabinovich, M. I. (1992). "Nonlinearities in Action," Springer-Verlag, New York.

Guckenheimer, J., and Holmes, P. J. (1983). "Nonlinear Oscillations; Dynamical Systems and Bifurcations of Vector Fields." Springer-Verlag, New York.

Jackson, A. (1989). "Perspectives in Nonlinear Dynamics," Vol. 1. Cambridge Univ. Press, New York.

Kapitaniak, T. (1996). "Controlling Chaos," Academic Press, London.

Lichtenberg, A. J., and Liebermann, M. A. (1983). "Regular and Stochastic Motion," Springer-Verlag, New York.

Minorsky, N. (1962). "Nonlinear Oscillations," Van Nostrand, Princeton, NJ.

Moon, F. C. (1992). "Chaotic and Fractal Dynamics," Wiley, New York.

Nayfeh, A. H., and Balachandran, B. (1993). "Nonlinear Dynamics," Wiley, New York.

Schuster, H. G. (1984). "Deterministic Chaos," Physik-Verlag GmbH, Weinheim, Federal Republic of Germany.

Strogatz, S. H. (1994). "Nonlinear Dynamics and Chaos," Addison Wesley, Reading, MA.

Toda, M. (1989). "Theory of Nonlinear Lattices," Springer-Verlag, Berlin.



Polarization and Polarimetry

Kent Rochford
National Institute of Standards and Technology

I. Polarization States
II. Polarizers
III. Retarders
IV. Mathematical Representations
V. Polarimetry

GLOSSARY

Birefringence The property of optically anisotropic materials, such as crystals, of having the phase velocity of propagation dependent on the direction of propagation and polarization. Numerically, birefringence is the refractive index difference between eigenpolarizations.

Diattenuation The property of having optical transmittance depend on the incident polarization state. In diattenuators, the eigenpolarizations will have principal transmittances $T_{\max}$ and $T_{\min}$, and diattenuation is quantified as $(T_{\max} - T_{\min})/(T_{\max} + T_{\min})$. Diattenuation may occur during propagation when absorption coefficients depend on polarization (also called dichroism) or at interfaces.

Eigenpolarization A polarization state that propagates unchanged through optically anisotropic materials. Eigenpolarizations are orthogonal in homogeneous polarization elements.

Jones calculus A mathematical treatment for describing fully polarized light. Light is represented by 2×1 complex Jones vectors and polarization components as 2×2 complex Jones matrices.

Mueller calculus A mathematical treatment for describing completely, partially, or unpolarized light. Light is represented by the 4×1 real Stokes vector and polarization components as 4×4 real Mueller matrices.

Polarimetry The measurement of the polarization state of light or the polarization properties (retardance, diattenuation, and depolarization) of materials.

Polarized light A light wave whose electric field vector traces a generally elliptical path. Linear and circular polarizations are special cases of elliptical polarization. In general, light is partially polarized, and is a mixture of polarized light and unpolarized light.

Polarizer A device with diattenuation approaching 1 that transmits one unique polarization state regardless of incident polarization.

Retardance The optical phase shift between two eigenpolarizations.

Unpolarized light Light of finite spectral width whose instantaneous polarization randomly varies over all states during the detection time. Not strictly a polarization state of light.

THE POLARIZATION state is one of the fundamental characteristics (along with intensity, wavelength,

and coherence) required to describe light. The earliest recorded observation of polarization effects was reported by Bartholinus, who observed double refraction in calcite in 1669. Huygens demonstrated the concept of polarization by passing light through two calcite crystals in 1690. Today, the measurement, manipulation, and control of polarization plays an important role in optical sciences.

I. POLARIZATION STATES

Light can be represented as an electromagnetic wave that satisfies Maxwell's equations. A transverse electromagnetic wave has electric and magnetic field components that are orthogonal to the direction of propagation. As the wave propagates, the strengths of these transverse fields oscillate in space and time, and the polarization state is defined by the direction of the electric field vector **E**.

For our discussion, we will use a right-handed Cartesian coordinate system with orthogonal unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$. A monochromatic plane wave $\mathbf{E}(z,t)$ traveling in vacuum along the $\hat{\mathbf{z}}$ direction with time t can be written as

$$\mathbf{E}(z,t) = \mathrm{Re}\{\hat{\mathbf{x}}E_x \exp[i(\omega t - k_0 z + \phi_x)] + \hat{\mathbf{y}}E_y \exp[i(\omega t - k_0 z + \phi_y)]\} \quad (1a)$$

or

$$\mathbf{E}(z,t) = \hat{\mathbf{x}}E_x \cos(\omega t - k_0 z + \phi_x) + \hat{\mathbf{y}}E_y \cos(\omega t - k_0 z + \phi_y), \quad (1b)$$

where ω is the angular optical frequency and E_x and E_y are the electric field amplitudes along the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ axes, respectively. The free-space wavenumber is $k_0 = 2\pi/\lambda$ for wavelength λ, and ϕ_x and ϕ_y are absolute phases. The difference in phase between the two component fields is then $\Delta\phi = \phi_y - \phi_x$. The direction of **E** and the polarization of the wave depend on the field amplitudes E_x and E_y and the phases ϕ_x and ϕ_y.

A. Linear Polarization

A wave is linearly polarized if an observer looking along the propagation axis sees the tip of the oscillating electric field vector confined to a straight line. Figure 1 depicts the wave propagation for two different linear polarizations when Eq. (1b) is plotted for $\phi_x = \phi_y = 0$. In Fig. 1, $E_y = 0$ and light is linearly polarized along the x axis; in the other example, light is polarized along the y axis when $E_x = 0$.

For a field represented by Eqs. (1a) and (1b), light will be linearly polarized whenever $\Delta\phi = m\pi$, where m is an integer; the direction of linear polarization depends on the magnitudes of E_x and E_y. For example, if $E_x = E_y$, the vector sum of these orthogonal fields yields a wave polarized at 45° from the x axis. If $E_x = -E_y$ (or if $E_x = E_y$ and $\Delta\phi = \pi$), the light is linearly polarized at −45°. For in-phase component fields ($\Delta\phi = 0$), the linear polarization is oriented at an angle $\alpha = \tan^{-1}(E_y/E_x)$ with respect to the x axis.

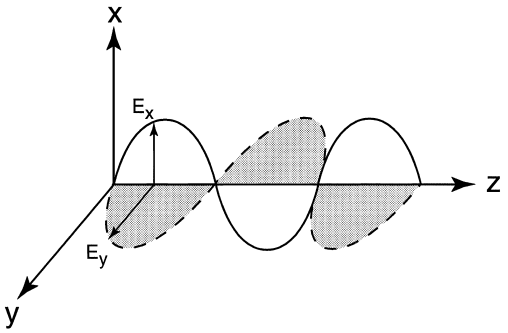


FIGURE 1 Two linear polarized waves. The electric field vector of *x*-polarized light oscillates in the *xz* plane. The shaded wave is *y*-polarized light in the *yz* plane.

In general, linear polarization states are often defined by an orientation angle, though descriptive terms such as x- or y-polarized, or vertical or horizontal, may be used. However, when a wave is incident upon a boundary two specific linearly polarized states are defined. The *plane of incidence* (Fig. 2) is the plane containing the incident ray and the boundary normal. The linear polarization in the plane of incidence is called p-polarization and the field component perpendicular to the plane is s-polarized. This convention is used with the Fresnel equations (Section II.A) to determine the transmittance, reflectance, and phase shift when light encounters a boundary.

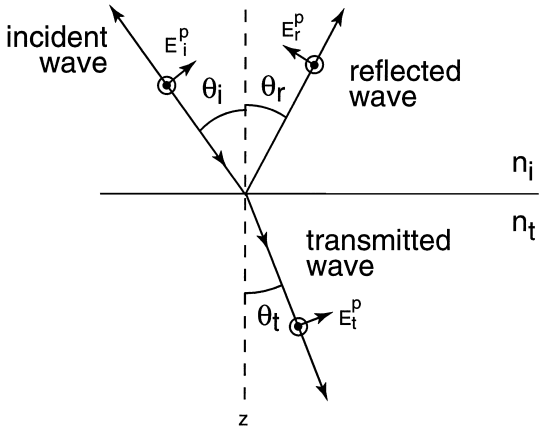


FIGURE 2 Light waves at a boundary. The plane of incidence coincides with the plane of the page. Incident, reflected, and transmitted *p*-polarized waves are in the plane of incidence. The corresponding *s*-polarizations (not shown) would be perpendicular to the plane of incidence.

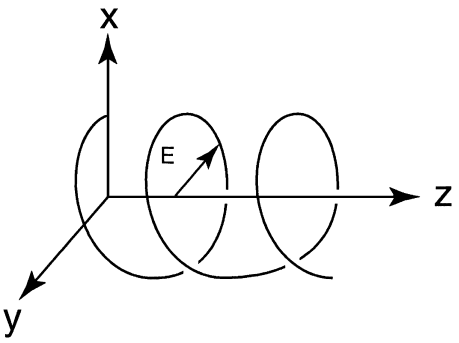


FIGURE 3 The electric field propagation for right-circular polarization, Eq. (2), when $t = 0$. At a fixed time, the tip of the electric field vector traces a right-handed corkscrew as the wave propagates along the $+z$ direction.

B. Circular Polarization

Another special case occurs when $E_x = E_y = E_0$ and the field components have a 90° relative phase difference $[\Delta\phi = (m + 1/2)\pi]$. If $\Delta\phi = \pi/2$, Eq. (1b) becomes

$$\begin{aligned}\mathbf{E}^{\text{rcp}} &= E_0[\hat{\mathbf{x}}\cos(\omega t - k_0 z) + \hat{\mathbf{y}}\cos(\omega t - k_0 z + \pi/2)] \\ &= E_0[\hat{\mathbf{x}}\cos(\omega t - k_0 z) - \hat{\mathbf{y}}\sin(\omega t - k_0 z)]. \qquad (2)\end{aligned}$$

As the wave advances through space the magnitude of $\mathbf{E}^{\text{rcp}}$ is constant but the tip of this electric field vector traces a circular path about the propagation axis at a frequency ω. A wave with this behavior is said to be right-circularly polarized.

Figure 3 shows the electric field vector for right-circular polarization when viewed at a fixed time ($t = 0$); here the field will trace a right-handed spiral in space. An observer looking toward the origin from a distant point ($z > 0$) would see the vector tip rotating counterclockwise as the field travels along z. In contrast, the same observer looking at a right-circularly polarized field at a *fixed* position (for example, $z = 0$) would see the vector rotation trace out a clockwise circle in the xy plane as time advances. This difference in the sense of rotation between space and time is often a source of confusion, and depends on notation (see Section I.F).

When light is left-circularly polarized the field traces out a left-handed spiral in space at a fixed time and a counterclockwise circle in time at a fixed position. Equation (1b) describes left-circular polarization when $E_x = E_y = E_0$ and $\Delta\phi = -\pi/2$:

$$\begin{aligned}\mathbf{E}^{\text{lcp}} &= E_0[\hat{\mathbf{x}}\cos(\omega t - k_0 z) + \hat{\mathbf{y}}\cos(\omega t - k_0 z - \pi/2)] \\ &= E_0[\hat{\mathbf{x}}\cos(\omega t - k_0 z) + \hat{\mathbf{y}}\sin(\omega t - k_0 z)]. \qquad (3)\end{aligned}$$

Right- and left-circular polarizations are orthogonal states and can be used as a basis pair for representing other polarization states, much as orthogonal linear states are combined to create circular polarization. Adding equal amounts of right- and left-circularly polarized light will yield a linearly polarized state. For example,

$$\tfrac{1}{2}\mathbf{E}^{\text{rcp}} + \tfrac{1}{2}\mathbf{E}^{\text{lcp}} = \hat{\mathbf{x}}E_0\cos(\omega t - k_0 z). \qquad (4)$$

In contrast, adding equal quantities of left- and right-circular polarization that are out of phase [by adding an additional π phase to both component fields in Eq. (2)] yields

$$-\tfrac{1}{2}\mathbf{E}^{\text{rcp}} + \tfrac{1}{2}\mathbf{E}^{\text{lcp}} = \hat{\mathbf{y}}E_0\sin(\omega t - k_0 z). \qquad (5)$$

In general, equal amounts of left- and right-circular polarization combine to produce a linear polarization with an azimuthal angle equal to half the phase difference.

C. Elliptical Polarization

For elliptically polarized light the electric field vector rotates at ω but varies in amplitude so that the tip traces out an ellipse in time at a fixed position z. Elliptical polarization is the most general state and linear and circular polarizations are simply special degenerate forms of elliptically polarized light. Because of this generality, attributes of this state can be applied to all polarization states.

The polarization ellipse (Fig. 4) can provide useful quantities for describing the polarization state. The azimuthal angle α of the semi-major ellipse axis from the x axis is given by

$$\tan(2\alpha) = \tan(\beta)\cos(\Delta\phi), \qquad (6)$$

where $\tan(\beta) = E_y/E_x$ and $0 \le \beta \le \pi/2$. The ellipticity $\tan|\varepsilon| = b/a$, the ratio of the semi-minor and semi-major axes, is calculated from the amplitudes and phases of Eq. (1) as

$$\tan(\epsilon) = \tan[\sin^{-1}(\sin 2\beta \sin\Delta\phi)/2]. \qquad (7)$$

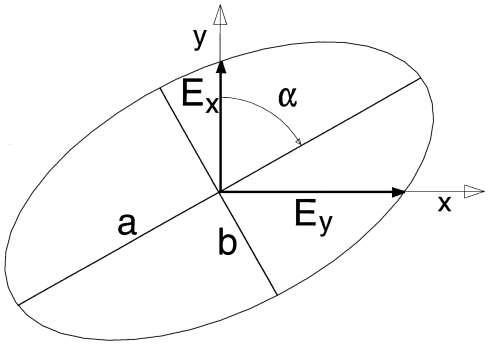


FIGURE 4 The polarization ellipse showing fields $\mathbf{E}_x$ and $\mathbf{E}_y$, ellipticity $\tan|\epsilon| = b/a$, and azimuthal angle α. The tip of the electric field **E** traces this elliptical path in the transverse plane as the field propagates down the z axis.

Polarization is right-elliptical when $0^\circ < \Delta\phi < 180^\circ$ and $\tan(\epsilon) > 0^\circ$ and left-elliptical when $-180^\circ < \Delta\phi < 0^\circ$ and $\tan(\epsilon) < 0^\circ$.

D. Unpolarized Light

Monochromatic, or single-frequency, light must necessarily be in some polarization state. Light that contains a band of wavelengths does not share this requirement.

Quasi-monochromatic light can be represented by modifying Eq. (1b) as

$$\mathbf{E}(z,t) = \mathrm{Re}(\hat{\mathbf{x}}E_x(t)\exp\{i[\omega_m t + \phi_x(t)]\} + \hat{\mathbf{y}}E_y(t)\exp\{i[\omega_m t + \phi_y(t)]\}) \quad (8)$$

where ω_m is the mean frequency of an electric field with bandwidth $\Delta\omega < \omega_m$. Taking the real part of this complex analytic representation yields the true field. Whereas the field amplitudes $E_i(t)$ and phases $\phi_i(t)$ are constants for strictly monochromatic light, these quantities fluctuate irregularly when the light has finite bandwidth. The pairs of functions $E_i(t)$ and $\phi_i(t)$ have statistical correlations that depend on the spectral bandwidth of the light source. The coherence time $\tau \sim 2\pi/\Delta\omega$ describes the time scale during which the pairs of functions show similar time response. For some brief time $t \ll \tau$, $E_i(t)$ and $\phi_i(t)$ are essentially constant, and $\mathbf{E}(t)$ possesses some elliptical polarization state, but a later field $\mathbf{E}(t+\tau)$ will have a different elliptical polarization. Light is described as unpolarized, or natural, if the time evolutions of the pairs of functions are totally uncorrelated within the detection time, and any polarization state is equally likely during these successive time intervals.

While strictly monochromatic light cannot be unpolarized, natural light can be polarized into any desired elliptical state by passing it through the appropriate polarizer. Indeed, when unpolarized light is incident on a polarizer, the detected output intensity is independent of the polarization state transmitted by the polarizer. This occurs because a unique polarization exists for an infinitesimal time $t \ll \tau$ and the average projection of these arbitrary states on a given polarizer is $\frac{1}{2}$ over the relatively long integration time of the detector. In the absence of dispersive effects, unpolarized light, when totally polarized by an ideal polarizer, will behave much like monochromatic polarized light.

It is often desirable to have unpolarized light, especially when the undesired polarization dependence of components degrades optical system performance. For example, the responsivity of photodetectors can exhibit polarization dependence and cause measurements of optical power to vary with the polarization even when intensity is constant. In some cases, pseudo-depolarizers are useful for modifying polarization to produce light that approximates unpolarized light (Section III.F). For quasi-monochromatic light, the orthogonal field components can be differentially delayed, or retarded, longer than τ, so that the fields become uncorrelated. Alternatively, repeatedly varying the polarization state over a time shorter than the detector response causes the measurement to include the influence of many polarization states. This method, known as polarization scrambling, can reduce some undesirable polarization effects by averaging polarizations.

The previous discussion implicitly assumes that the light has uniform properties over the wavefront. However, the polarization can be varied over the spatial extent of the beam using a spatially varying retardance. Further description of these methods and their limitations is found in the discussion on optical retarders.

E. Degree of Polarization

Light that is neither polarized nor unpolarized is partially polarized. The fraction of the intensity that is polarized for a time much longer than the optical period is called the degree of polarization P and ranges from $P = 0$ for unpolarized light to $P = 1$ when a light beam is completely polarized in any elliptical state. Light is partially polarized when $0 < P < 1$. Partially polarized light occurs when $E_i(t)$ and $\phi_i(t)$ are not completely uncorrelated, and the instantaneous polarization states are limited to a subset of possible states. Partially polarized light may also be represented as a sum of completely polarized and unpolarized components.

We can also define a degree of linear polarization (the fraction of light intensity that is linearly polarized) or a degree of circular polarization (the fraction that is circularly polarized). Degrees of polarization can be described formally using the coherency matrix or Stokes vector formalism described in Section IV.

F. Notation

The choice of coordinate system and the form of the field in Eqs. (1a) and (1b) is not unique. We have chosen a right-handed coordinate system such that the cross-product is $\hat{\mathbf{x}} \times \hat{\mathbf{y}} = \hat{\mathbf{z}}$ and used fields with a time dependence $\exp[i(\omega t - kz)]$ rather than the complex conjugate $\exp[-i(\omega t - kz)]$. Both choices are equally valid, but may result in different descriptions of the same polarization states. Descriptions of circular polarization in particular are often contradictory because of the confusion arising from the use of varied conventions. In this article we follow the "Nebraska Convention" adopted in 1968 by the participants of the Conference on Ellipsometry at the University of Nebraska.

Also, the choice of the Cartesian basis set for describing the electric field is common but not obligatory. Any polarization state can be decomposed into a combination of any pair of orthogonal polarizations. Thus Eqs. (1a) and (1b) could be written in terms of right- and left-circular states or orthogonal elliptical states.

II. POLARIZERS

An ideal polarizer transmits only one unique state of polarization regardless of the state of the incident light. Polarizers may be delineated as linear, circular, or elliptical, depending on the state that is produced. Linear polarizers that transmit a linear state are the most common and are often simply called "polarizers."

The transmission axis of a linear polarizer corresponds to the direction of the output light's electric field oscillation. This axis is fixed by the device, though polarizers can be oriented (rotated normal to the incident light) to select the azimuthal orientation of the output state. When linearly polarized light is incident on a linear polarizer, the transmittance T from the polarizer follows Malus's law,

$$T = \cos^2\theta, \tag{9}$$

where θ is the angle between the input polarization's azimuth and the polarizer's transmission axis. When the incident light is formed by a linear polarizer, Eq. (9) describes the transmission through two polarizers with angle θ between transmission axes. In this configuration the second polarizer is often called an analyzer, and the polarizer and analyzer are said to be *crossed* when the transmittance is minimized ($\theta = 90^\circ$).

Since an ideal polarizer transmits only one polarization state it must block all others. In practice polarizers are not ideal, and imperfect polarizers do not exclude all other states. For an imperfect polarizer Malus's law becomes

$$T = (T_{\max} - T_{\min})\cos^2\theta + T_{\min}, \tag{10}$$

where $T_{\max}$ and $T_{\min}$ are called the principal transmittances, and transmittance T varies between these values. The *extinction ratio* $T_{\min}/T_{\max}$ provides a useful measure of polarizer performance. Diattenuation is the dependence of transmittance on incident polarization, and can be quantified as $(T_{\max} - T_{\min})/(T_{\max} + T_{\min})$, where the maximum and minimum transmittances occur for orthogonal polarizations in homogeneous elements. (Homogeneous polarization elements have eigenpolarizations that are orthogonal and we consider such elements exclusively in this article.) Polarizers are optical elements that have a diattenuation approaching 1.

Most interfaces with nonnormal optical incidence will exhibit some linear diattenuation since the Fresnel reflection and transmission coefficients depend on the polarization. High-performance polarizers exploit these effects to achieve very high diattenuations by differentially reflecting and transmitting orthogonal polarizations. In contrast, dichroism is a material property in which diattenuation occurs as light travels through the medium. Most commercial polarizers exploit dichroism, polarization-dependent reflection or refraction in birefringent crystals, or polarization-dependent reflectance and transmittance in dielectric thin-film structures.

A. Fresnel Equations

Maxwell's equations applied to a plane wave at an interface between two dielectric media provide the relationship among incident, transmitted, and reflected wave amplitudes and phases. Figure 2 shows the electric fields and wavevectors for a wave incident upon the interface between two lossless, isotropic dielectric media. The *plane of incidence* contains all three wavevectors and is used to define two specific linear polarization states; p-polarized light has its electric field vector within the plane of incidence, and s-polarized light is perpendicular to this plane. The law of reflection, $\theta_i = \theta_r$, provides the direction of the reflected wave. The refraction angle is given by Snell's law,

$$n_i \sin\theta_i = n_t \sin\theta_t. \tag{11}$$

Fresnel's equations yield the amplitudes of the transmitted field E_t and reflected field E_r as fractions of the incident field E_i. For p-polarized light in isotropic, homogeneous, dielectric media, the amplitude reflectance r_p is

$$r_p = \left(\frac{E_r}{E_i}\right)_p = \frac{n_t\cos\theta_i - n_i\cos\theta_t}{n_i\cos\theta_t + n_t\cos\theta_i} \tag{12}$$

and amplitude transmittance t_p is

$$t_p = \left(\frac{E_t}{E_i}\right)_p = \frac{2n_i\cos\theta_i}{n_i\cos\theta_t + n_t\cos\theta_i}. \tag{13}$$

For s-polarized light, the corresponding Fresnel equations are

$$r_s = \left(\frac{E_r}{E_i}\right)_s = \frac{n_i\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t} \tag{14}$$

and

$$t_s = \left(\frac{E_t}{E_i}\right)_s = \frac{2n_i\cos\theta_i}{n_i\cos\theta_i + n_t\cos\theta_t}. \tag{15}$$

The Fresnel reflectance for cases $n_i/n_t = 1.5$ and $n_t/n_i = 1.5$ is shown in Fig. 5. At an incidence angle $\theta_B = \tan^{-1}(n_t/n_i)$ (for $n_t > n_i$), known as the Brewster angle, $r_p = 0$ and p-polarized light is totally transmitted. In a *pile-of-plates* polarizer, plates of glass are oriented at the

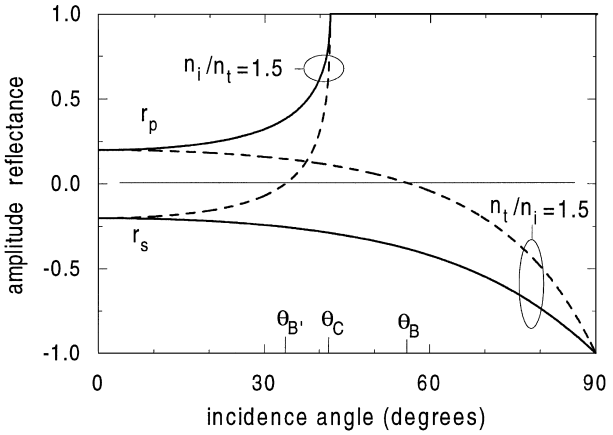


FIGURE 5 Fresnel reflectances for p-polarized (solid curve) and s-polarized (dashed) light for cases $n_i/n_t = 1.5$ and $n_t/n_i = 1.5$. The amplitude reflectance is 0 for p-polarized light at the Brewster angle θ_B, and is one for all polarizations when the incidence angle is $\theta \geq \theta_C$.

Brewster angle so that only s-polarized light is reflected from each plate, and the successive diattenuations from each plate increase the degree of polarization of transmitted light.

When $n_i > n_t$, both polarizations may be completely reflected if the incidence angle is larger than the critical angle θ_c,

$$\theta_c = \sin^{-1} \frac{n_t}{n_i}. \tag{16}$$

When $\theta_i \geq \theta_c$, the light undergoes *total internal reflection* (TIR). For these incidence angles no net energy is transmitted beyond the interface and an *evanescent* field propagates along the direction θ_t. The reflectance can be reduced from 1 if the medium beyond the interface is thinner than a few wavelengths and followed by a higher refractive index material. The resulting *frustrated total internal reflection* allows energy to flow across the interface, leading to nonzero transmittance. For this reason, TIR devices using glass–air interfaces must be kept free of contaminants that may frustrate the TIR. Birefringent crystal polarizers obtain very high extinction ratios by transmitting one linear polarization while forcing the orthogonal polarization to undergo TIR.

B. Birefringent Crystal Polarizers

Birefringent polarizers spatially separate an incident beam into two orthogonally polarized beams. In a conventional polarizer, the undesired polarization is eliminated by directing one beam into an optical absorber so that a single polarization is transmitted. Alternatively, a polarizing beamsplitter transmits two distinct orthogonally polarized beams that are angularly separated or displaced.

In birefringent materials, the incident polarization is decomposed into two orthogonal states called principal polarizations or eigenpolarizations. When the eigenpolarizations travel at the same velocity (and see the same refractive index), the direction of propagation is called an *optic axis* (see Section III.A). When light does not travel along an optic axis, the eigenpolarizations see different refractive indices and thus propagate at different velocities through the material.

When light enters or exits a birefringent material at a nonnormal angle θ that is not along an optic axis, the eigenpolarizations refract at different angles, undergoing what is termed double refraction. Also, each eigenpolarization may encounter different reflectance or transmittance at interfaces (since Fresnel coefficients depend on the refractive indices), and diattenuation results. Complete diattenuation occurs if one eigenpolarization undergoes total internal reflection while the other eigenpolarization is transmitted.

Most birefringent polarizers are made from calcite, a naturally occurring mineral. Calcite is abundant in its polycrystalline form, but optical-grade calcite required for polarizers is rare, which makes birefringent polarizers more costly than most other types. Calcite transmits from below 250 nm to above 2 μm and is used for visible and near-infrared applications. Other birefringent crystals, such as magnesium fluoride (with transmittance from 140 nm to 7 μm), can be used at some wavelengths for which calcite is opaque.

Prism polarizers are composed of two birefringent prisms cut at an internal incidence angle that transmits only one eigenpolarization while totally internally reflecting the other (Fig. 6). The prisms are held together by a thin cement layer or may be separated by an air gap and externally held in place for use with higher power laser beams. The transmitted beam contains only one eigenpolarization since the orthogonal polarization is completely reflected. The prisms are aligned with parallel optic axes, so that this transmitted beam undergoes very small deviations, usually less than 5 min of arc. Often the reflected beam also

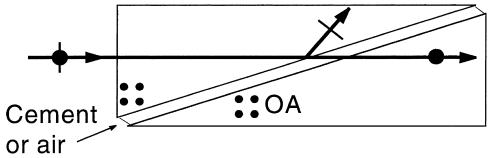


FIGURE 6 Glan–Thompson prism polarizer. At the interface, p-polarized light reflects (and is typically absorbed by a coating at the side of the prism) and s-polarized light is transmitted. The optic axes (shown as dots) are perpendicular to the page.

contains a small amount of the transmitted eigenpolarization since nonzero reflectance results if the refractive indices of the cement and transmitted eigenpolarization are not exactly equal. Because the reflected beam has poorer extinction, it is usually eliminated by placing an index-matched absorbing layer on the side face toward which light is reflected.

Glan prism polarizers are the most common birefringent crystal polarizer. They exhibit superior extinction; extinction ratios of 10^{-5}–10^{-6} are typical, and extinctions below 10^{-7} are possible. The small residual transmittance can arise from material imperfection, scattering at the prism faces, or misalignment of the optic axes in each prism of the polarizer.

Because total internal reflection requires incidence angles larger than θ_c, the polarizer operates over a limited range of input angles that is often asymmetric about normal incidence. The semi-field angle is the maximum angle for which output light is completely polarized regardless of the rotational orientation of the polarizer (that is, for any azimuthal angle of output polarization). The field angle is twice the semi-field angle. The field angle depends on the refractive index of the intermediate layer (cement or air) and the internal angle of the contacted prisms. Since the incidence angle at the contacting interface depends in part on the refractive index when light is nonnormally incident on the polarizer, the field angle is wavelength dependent.

Birefringent crystal polarizing beamsplitters transmit two orthogonal polarizations. Glan prism polarizers can act as beamsplitters if the reflected beam exits through a polished surface, though extinction is degraded. Polarizing beamsplitters with better extinction separate the beams through refraction at the interface. In Rochon prisms, light linearly polarized in the plane normal to the prism is transmitted undeviated, while the orthogonal polarization is deviated by an angle dependent on the prism wedge angle and birefringence (Fig. 7a). Sénarmont polarizing beam splitters are similar, but the polarizations of the deviated and undeviated beams are interchanged. Wollaston polarizers (Fig. 7b) deviate both output eigenpolarizations with nearly equal but opposite angles when the input beam is normally incident. For all these polarizers, the deviation angle depends on the wedge angle and varies with wavelength.

C. Interference Polarizers

The Fresnel equations show that the transmittance and reflectance of obliquely incident light will depend on the polarization. Dielectric stacks made of alternating high- and low-refractive index layers with quarter-wave optical thickness can be tailored to provide reflectances and transmittances with large diattenuation. Optical thickness depends on incidence angle, and polarizers based on quarter-wave layers are sensitive to incidence angle and wavelength. Designs that increase the wavelength range do so at the expense of input angle range, and vice versa. Polarizing beamsplitter cubes are made by depositing the stack on the hypotenuse of a right-angle prism and cementing the coated side to the hypotenuse of a second prism.

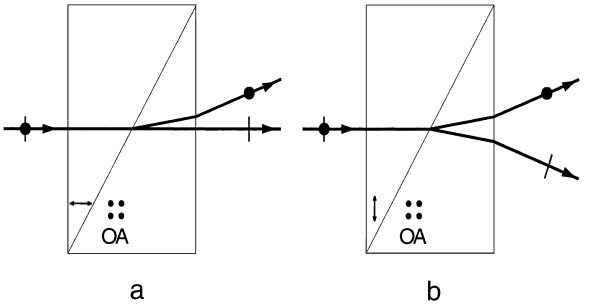


FIGURE 7 (a) Rochon and (b) Wollaston polarizers. The directions of the optic axes are shown in each prism (as dots for axes perpendicular to page and as a two-arrow line for axes in the plane of the page).

The extinction of these devices is limited by the defects in the coating layers or the optical quality of the optical substrate material through which light must pass. The state of polarization may also be altered by the birefringence in the substrate. Commercial thin-film polarizers are available with an extinction of about 10^{-5}.

D. Dichroic Polarizers

Some molecules are optically anisotropic, and light polarized along one molecular direction may undergo greater absorption than perpendicularly polarized light. When these molecules are randomly oriented, this molecular-level diattenuation will average out as the light propagates through the thickness, and bulk diattenuation may not be observed. However, linear polarizers can be made by orienting dichroic molecules or crystals in a plastic or glass matrix that maintains a desired alignment of the transmission axes. Extinction ratios between 10^{-2} and 10^{-5} are possible in oriented dichroics in the visible and near-infrared regions.

Dichroic sheet polarizers are available with larger areas and at lower cost than other polarizer types. Also, the acceptance angle, or maximum input angle from normal incidence that does not result in degraded extinction, is typically large in dichroics because diattenuation occurs during bulk propagation rather than at interfaces. However, the maximum transmittance of these polarizers may be significantly less than unity since the transmission axis

may also absorb light. Because absorbed light will heat the material and may cause damage at high power, incident powers are limited.

III. RETARDERS

Retarders are devices that induce a phase difference, or retardation, between orthogonally polarized components of a light wave. Linear retarders are the most common and produce a retardance $\Delta\phi = \phi_y - \phi_x$ [using the notation of Eqs. (1a) and (1b)] between orthogonal linear polarizations. Circular retarders cause a phase shift between right- and left-circular polarizations and are often called rotators because circular retardance changes the azimuthal angle of linearly polarized light. Because the polarization state of light is determined by the relative amplitudes and phase shifts between orthogonal components, retarders are useful for altering and controlling a wave's polarization. In fact, an arbitrary polarization state can be converted to any other state using an appropriate retarder.

A. Linear Birefringence

In optically anisotropic materials, such as crystals, the phase velocity of propagation generally depends on the direction of propagation and polarization. The *optic* axes are propagation directions for which the phase velocity is independent of the azimuth of linear polarization. For other propagation directions, two orthogonal eigenaxes perpendicular to the propagation define the linear polarizations of waves that propagate through the crystal with constant phase velocity. These *eigenpolarizations* are linear states whose refractive indices are determined by the crystal's dielectric tensor and propagation direction. Light polarized in an eigenpolarization will propagate through an optically anisotropic material with unchanging polarization, while light in other polarization states will change with distance as the beam propagates.

Uniaxial crystals and materials that behave uniaxially are commonly used in birefringent retarders and polarizers. These crystals have a single optic axis, two principal refractive indices n_o and n_e, and a linear birefringence $\Delta n = n_e - n_o$. When light travels parallel to the optic axis, the eigenpolarizations are degenerate, and all polarizations propagate with index n_o. For light traveling in other directions, one eigenpolarization has refractive index n_o and the other's varies with direction between n_o and n_e (and equals n_e when the propagation is perpendicular to the optic axis).

B. Waveplates

Waveplates are linear retarders made from birefringent materials. Rewriting Eq. (1a) for propagation through a birefringent medium of length L yields

$$\mathbf{E}(z = L, t) = \mathrm{Re}\{E_x \exp[i(\omega t - k_0 n_x L)] + E_y \exp[i(\omega t - k_0 n_y L)]\}, \quad (17)$$

where the x and y directions coincide with eigenpolarizations and the absolute phases are initially equal (at $z = 0$, $\phi_x = \phi_y = 0$). The retardance $\Delta\phi = k_0(n_x - n_y)L$ is the relative phase shift between eigenpolarizations and depends on the wavelength, the propagation distance, and the difference between the refractive indices of the eigenpolarizations. If the z axis is an optic axis, then $n_x = n_y = n_o$, and there is no retardance; if $\hat{\mathbf{z}}$ is perpendicular to an optic axis, the retardance is $\Delta\phi = \pm k_0(n_o - n_e)L$. In general, the retardance over a path of length L in a material with birefringence Δn is given by

$$\Delta\phi = 2\pi\,\Delta n L/\lambda. \quad (18)$$

Retardance may be specified in radians, degrees [$\Delta\phi = 360^\circ \cdot (n_o - n_e)L/\lambda_0$], or length [$\Delta\phi = (n_o - n_e)L$].

A waveplate that introduces a π-radian or 180° phase shift between the eigenpolarizations is called a *half-wave plate*. Upon exiting the plate, the two eigenpolarizations have a $\lambda/2$ relative delay and are exactly out of phase. A half-wave plate requires a birefringent material with thickness given by

$$L_{\lambda/2} = \frac{\lambda_0(2m + 1)}{2\,|n_o - n_e|}, \quad (19)$$

where the waveplate order m is a positive integer that need not equal 0 since additional retardances of 360° do not affect the phase relationship. *Quarter-wave plates* are another common component and provide phase shifts of 90° or $\pi/2$.

The eigenaxis with the lower refractive index (n_o in positive uniaxial crystals such as quartz, and n_e in negative uniaxial crystals such as calcite) is called the fast axis of the retarder due to the faster phase velocity and is often marked by the manufacturer. The eigenaxes can be identified by rotating the retarder between crossed polarizers until the transmittance is minimized. When the polarizer transmission axis coincides with the retarder eigenaxis, the input polarization matches the eigenpolarization, and the light travels through the crystal unchanged until blocked by the analyzer. An input different from the eigenpolarization will exit the crystal in a different polarization state and will not be completely blocked by the analyzer.

Waveplates are commonly made using quartz, mica, or plastic sheets that are stretched to produce an anisotropy that gives rise to birefringence. At visible wavelengths, $n_e - n_o \sim 0.009$ for quartz, and the corresponding zeroth-order ($m = 0$) quarter-wave plate thickness of ~ 40 μm

poses a severe manufacturing challenge. Mica can be cleaved into thin sections to obtain zeroth-order retardance, but the resulting waveplate usually has poorer spatial uniformity. Polymeric materials often have lower birefringence and can be most easily fabricated into zeroth-order waveplates.

In many applications, retardance of integral multiples of 2π is unimportant, and *multiple-order* ($m \geq 0$) waveplates are often lower in cost because the increased thickness eases fabrication. However, this approach can result in increased retardance errors. For example, retardance depends on the wavelength [explicitly in Eq. (18) or through dispersion]. Also, retardance can change with temperature or with nonnormal incidence angles that vary the optical thickness and propagation direction. Retardance errors arising from changes in wavelength, temperature, or incidence angle linearly increase with thickness and make multiple-order waveplates unadvisable in applications that demand accurate retardance.

Compound zeroth-order waveplates represent a compromise between manufacturability and performance when true zeroth-order waveplates are not easily obtained. When two similar waveplates are aligned with orthogonal optic axes, the phase shifts in each waveplate have opposite sign and the combined retardance will be the difference between the two retardances. Compound zeroth-order retarders are made by combining two multiple-order waveplates in this way so that the net retardance is less than 2π. For example, two multiple-order waveplates with retardance $\Delta\phi_1 = 20\pi + \pi/2$ and $\Delta\phi_2 = -20\pi$ can be combined to yield a compound zeroth-order quarter-wave plate. Compound zeroth-order waveplates exhibit the same wavelength and temperature dependence as zeroth-order waveplates since retardance errors are proportional to the difference of plate thicknesses. However, input angle dependence is the same as in a multiple-order waveplate with equivalent total thickness.

C. Compensators

A compensator is a variable linear retarder that can be adjusted over a continuous range of values (Fig. 8). In a Babinet compensator, two wedged plates of birefringent material are oriented with their optic axes perpendicular. In this arrangement, the individual wedges impart opposite signs of retardance, and the net retardance is the difference between the individual magnitudes. The magnitudes depend on the thickness of each wedge traversed by the optical beam. Typically one wedge is fixed and the other translated by a micrometer drive so that this moving wedge presents a variable thickness in the beam path, and the net retardance depends on the micrometer adjustment. The use of two wedges eliminates

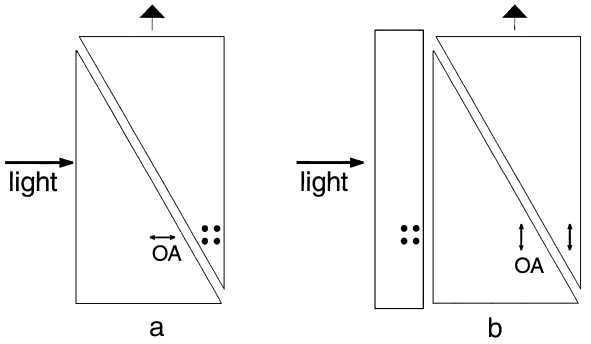


FIGURE 8 (a) Babinet and (b) Soleil–Babinet compensators. One wedge moves in the direction of the vertical arrow to adjust the retardance. The direction of the optic axes are shown using notation from Fig. 7.

the beam deviation and the output beam is collinear to the input.

The Babinet compensator has the disadvantage that the retardance varies across the optical beam because the relative thicknesses of each wedge and corresponding net retardance vary over the beam in the direction of wedge travel. This can be overcome using a Soleil (or Babinet–Soleil) compensator. In this device the two wedged pieces have coincident optic axes and translation of the moving wedge changes the total thickness and retardance of the combined retarder. The total thickness of this two-wedge piece is now constant over the useful aperture. A parallel plate of fixed retardance is placed after the wedge, in the same manner as a compound zeroth-order retarder, to improve performance.

D. Rhombs

Retarders can also be fabricated of materials that do not exhibit birefringence. The phase shift between s- and p-polarized waves that occurs at a total internal reflection (Section II, Fresnel equations) can be exploited to obtain a linear retarder. When light is incident at angles larger than the critical angle, the retardance at the reflection is

$$\Delta\phi = \phi_p - \phi_s = 2\tan^{-1}\left[\frac{\cos\theta_i\sqrt{\sin^2\theta_i - (n_i/n_t)^2}}{\sin^2\theta_i}\right] \tag{20}$$

and depends on the incidence angle and refractive indices. A *Fresnel rhomb* is a solid parallelogram fabricated so that a beam at normal incidence at the entrance face totally reflects twice within the rhomb to provide a net retardance of $\pi/2$. This retarder is, however, very sensitive to the incidence angle and laterally displaces the beam. Concatenating two Fresnel rhombs (Fig. 9) provides collinear output and can greatly reduce the sensitivity of retardance to incident angle since retardance changes at the first pair of reflections are partially canceled by the second pair.

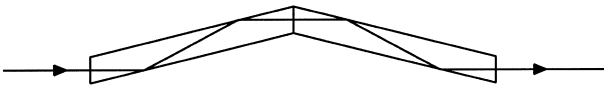


FIGURE 9 Two Fresnel rhombs concatenated to form a Fresnel double rhomb.

Total-internal-reflection retarders are less sensitive to wavelength variation than waveplates whose retardance increases with L/λ since the rhomb retardance does not depend on the optical path length. Wavelength dependence is limited only by the material dispersion $dn/d\lambda$, which contributes small retardance changes. Thus, rhomb devices are more nearly achromatic than waveplates and can be operated over ranges of 100 nm or more. Rhomb devices are much larger than waveplates, and the clear aperture has practical limits since increasing cross section requires a proportional increase in length. Performance can also be compromised by the presence of birefringence in the bulk glass. Birefringence, arising from stresses in material production or optical fabrication, can lead to spatial variations and path-length dependence, and limit retardance stability to several degrees if not mitigated.

E. Circular Retarders

Some materials can exhibit circular birefringence, or *optical activity*, in which the eigenpolarizations are right- and left-circular and the retardance is a phase shift between these two circular states. Circular retarders are often called rotators because incident linear polarization will generally exit at a different azimuthal angle that depends on the *rotary power* (circular retardance per unit length) and thickness. A material that rotates linearly polarized light clockwise (as viewed by an observer facing the light source) is termed dextrorotary or right-handed, while counterclockwise rotation occurs in levorotary, or left-handed, materials. The sense of rotation is fixed with respect to the propagation direction; if the beam exiting an optically active material is reflected back through the material, the polarization will be restored to the initial azimuth. Thus a double pass through an optically active material will cause no net rotation of linear polarization.

Crystalline quartz exhibits optical activity that is most evident when propagation is along the optic axis and retardance is absent. The property is not limited to crystalline materials, however; molecules that are chiral (that lack plane or centrosymmetry and are not superposable on their mirror image) can yield optical activity. *Enantiomers* are chiral molecules that share common molecular formulas and ordering of atoms but differ in the three-dimensional arrangement of atoms; separate enantiomers have equal rotary powers but differ in the sense of rotation. Liquids and solutions of chiral molecules such as sugars may be optically active if an excess of one enantiomer is present.

In solution, each enatiomeric form will rotate light, and the net rotation depends on the relative quantities of dextrorotary and levorotary enantiomers. Mixtures with equal quantities of enantiomers present are called racemic and the net rotation is zero. Most naturally synthesized organic chiral molecules, for example, sugars and carbohydrates, occur in only one enatiomeric form. Saccharimetry, the measurement of the optical rotary power of sugar solutions, is used to determine the concentration of sugar in single-enantiomer solutions.

F. Electrooptic and Magnetooptic Effects

In some materials, retardance can be induced by an electric or magnetic field. These effects are exploited to create active devices that produce an electrically controllable retardance.

Crystals that are not centrosymmetric may exhibit a linear birefringence proportional to an applied electric field called the linear electrooptic effect or Pockels effect. In these materials, applied fields cause an otherwise isotropic crystal to behave uniaxially (and uniaxial crystals to become biaxial). Crystal symmetry determines the direction of the optic axes and the form of the electrooptic tensor. The magnitude of the induced birefringence thus depends on the polarization direction, the applied field strength and direction, and the material.

The electrically induced birefringence can be appreciable in some materials, and the Pockels effect is widely used in retardance modulators, phase modulators, and amplitude modulators. Modulators are often characterized by their half-wave voltage V_π, or the voltage needed to cause a 180° phase shift or retardance. V_π can vary from ~10 V in waveguide modulators to hundreds or thousands of volts in bulk modulators.

The Kerr, or quadratic, electrooptic effect occurs in solids, liquids, or gases and has no symmetry requirements. In this effect, the linear birefringence magnitude is proportional to the square of the applied electric field and the induced optic axis is parallel to the field direction. The effect is typically smaller than the Pockels effect and is often negligible in Pockels materials.

The Faraday effect is an induced circular birefringence that is proportional to an applied magnetic field. It is often called Faraday rotation because the circular birefringence rotates linearly polarized light by an angle proportional to the field. The Faraday effect can occur in all materials, though the magnitude is decreased by birefringence.

In contrast to optical activity, the sense of Faraday rotation is determined by the direction of the magnetic field. Thus, a double-pass configuration in which light exiting a

Faraday rotator reflects and propagates back through the material will yield twice the rotation of a single pass. This property is exploited in optical isolators, or components that transmit light in only one direction. In the simplest isolators, a 45° Faraday rotator is placed between polarizers with transmission axes at 0° and 45°. In the forward direction, light linearly polarized at 0° is azimuthally rotated 45° to coincide with the analyzer axis and is fully transmitted; backward light input at 45° rotates to 90° and is completely blocked by the polarizer at 0°.

Faraday mirrors, made by combining a 45° Faraday rotator with a plane mirror, have the extraordinary property of "unwinding" polarization changes caused by propagation. Polarized light that passes through an arbitrary retarder, reflects off a Faraday mirror, and retraces the input path will exit with a fixed polarization for all magnitudes or orientations of the retarder so long as the retardance is unchanged during the round-trip time. When the input light is linearly polarized, the return light is always orthogonally polarized for all intervening retardances. These devices find applications in fiber optic systems since bend-induced retardance is difficult to control in an ordinary optical fiber.

G. Pseudo-Depolarizers

Conversion of a polarized, collimated light beam into a beam that is truly unpolarized is difficult. Methods for obtaining truly unpolarized light rely on diffuse scattering, such as passing light through ground glass plates or an integrating sphere. These methods result in light propagating over a large range of solid angles and decrease the irradiance, or power per unit area, away from the depolarizer. The loss is often unacceptable when a collimated beam is needed.

Approximations to the unpolarized state can be created using pseudo-depolarizers that produce a large variety of states over time, wavelength, or the beam cross section. As described in Section I, temporal decorrelation requires that the beam propagate through a retardance that is much larger than the light's coherence length $L_c = c\tau \approx 2\pi c/\Delta\omega$. If nonmonochromatic, linearly polarized light bisects the axes of a waveplate with sufficiently large retardance, the two linear eigenpolarizations will emerge with a relative phase shift that rapidly and arbitrarily changes on the order of the coherence time. At any moment the instantaneous output state will be restricted to a point on the Poincaré sphere (see Section IV) along the great circle connecting the ±45° and circular polarization states. When the detector is slower than τ, the averaged response will include the influence of all these states.

Lyot depolarizers are configurations of two retarders that perform this temporal decorrelation for any input polarization state. These are commonly made by concatenating thick birefringent plates that act as high-order waveplates or by connecting lengths of polarization-maintaining (PM) fiber. PM fiber has about one wavelength of retardance every few millimeters, and can be obtained in lengths sufficient to decorrelate multimode laser light.

A polarized light beam can also be converted to a beam with a spatial distribution of states to approximate unpolarized light, without the requirements on spectral bandwidth. For example, the retardance across a wedged waveplate is not spatially uniform, and an incident beam will exit with a spatially varying polarization. When detected by a single photodetector, the influence of all the states will be averaged in the output response. These methods often satisfy needs for unpolarized light, but clearly depend on the details and requirements of the application.

IV. MATHEMATICAL REPRESENTATIONS

Several methods have been developed to facilitate the representation of polarization states, polarization elements, and the evolution of polarization states as light passes through components. Using quasimonochromatic fields, the 2×2 coherency matrix can be used to represent polarizations and determine the degree of polarization of light. The four-element Stokes vector describes the state of light using readily measurable intensities and can be related to the coherency matrix. Mueller calculus represents optical components as real 4×4 matrices; when combined with Stokes vectors it provides a quantitative description of the interaction of light and optical components. In contrast, Jones calculus represents components using complex 2×2 matrices and represents light using two-element electric field vectors. Jones calculus cannot describe partially polarized or unpolarized light, but retains phase information so that coherent beams can be properly combined. Finally, the Poincaré sphere is a pictorial representation that is useful for conceptually understanding the interaction between retarders and polarization states. A brief discussion introduces each of these methods.

A. Coherency Matrix

Using Eq. (8), we can define orthogonal field components of a quasi-monochromatic plane wave $E_x = E_x(t)\exp[i(\omega t - k_0 z + \phi_x(t))]$ and likewise for E_y. The coherency matrix $\mathbf{J}$ is given by

$$\mathbf{J} = \begin{bmatrix} \langle E_x E_x^* \rangle & \langle E_x E_y^* \rangle \\ \langle E_y E_x^* \rangle & \langle E_y E_y^* \rangle \end{bmatrix} = \begin{bmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{bmatrix}, \qquad (21)$$

where the angle brackets denote a time average and the asterisk denotes the complex conjugate. The total irradiance I is given by the trace of the matrix, $\mathrm{Tr}(\mathbf{J}) = J_{xx} + J_{yy}$, and the degree of polarization is

$$P = \sqrt{1 - \frac{4|\mathbf{J}|}{(J_{xx} + J_{yy})^2}}, \tag{22}$$

where $|\mathbf{J}|$ is the determinant of the matrix. Recalling the notation for elliptical light, one can find the azimuthal angle α of the semi-major ellipse axis from the x axis and the ellipticity angle ϵ of the polarized component as

$$\begin{aligned} \alpha &= \frac{1}{2}\tan^{-1}\left[\frac{J_{xx} + J_{yy}}{J_{xx} - J_{yy}}\right] \\ \epsilon &= \frac{1}{2}\tan^{-1}\left[\frac{-i(J_{xy} - J_{yx})}{P(J_{xx} - J_{yy})}\right]. \end{aligned} \tag{23}$$

Partially polarized light can be decomposed into polarized and unpolarized components and expressed using coherency matrices as $\mathbf{J} = \mathbf{J}_\mathrm{p} + \mathbf{J}_\mathrm{u}$. Thus the state of the polarized portion of light can be extracted from the coherency matrix even when light is partially polarized. The coherency matrix representation of several states is provided in Table I.

B. Mueller Calculus

In Mueller calculus the polarization state of light is represented by a four-element Stokes vector $\mathbf{S}$. The Stokes parameters s_0, s_1, s_2, and s_3 are related to the coherency matrix elements or the quasi-monochromatic field representation through

$$\begin{aligned} s_0 &= J_{xx} + J_{yy} &&= \langle E_x(t)^2\rangle + \langle E_y(t)^2\rangle \\ s_1 &= J_{xx} - J_{yy} &&= \langle E_x(t)^2\rangle - \langle E_y(t)^2\rangle \\ s_2 &= J_{xy} + J_{yx} &&= 2\langle E_x(t)E_y(t)\cos(\Delta\phi)\rangle \\ s_3 &= i(J_{yx} - J_{xy}) &&= 2\langle E_x(t)E_y(t)\sin(\Delta\phi)\rangle, \end{aligned} \tag{24}$$

where the angle brackets denote a time averaging required for nonmonochromatic light. Each Stokes parameter is related to the difference between light intensities of specified orthogonal pairs of polarization states. Thus, the Stokes vector is easily found by measuring the power P_t transmitted through six different polarizers. Specifically,

$$S = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} P_{0^\circ} + P_{90^\circ} \\ P_{0^\circ} - P_{90^\circ} \\ P_{+45^\circ} - P_{-45^\circ} \\ P_\mathrm{rcp} - P_\mathrm{lcp} \end{bmatrix}, \tag{25}$$

so that s_0 is the total power or irradiance of the light beam, s_1 is the difference of the powers that pass through horizontal (along $\hat{\mathbf{x}}$) and vertical (along $\hat{\mathbf{y}}$) linear polarizers, s_2 is the difference between $+45^\circ$ and -45° linearly polarized powers, and s_3 is the difference between right- and left-circularly polarized powers. The values of the Stokes parameters are limited to $s_0^2 \geq s_1^2 + s_2^2 + s_3^2$ and are often normalized so that $s_0 = 1$ and $-1 \leq s_1, s_2, s_3 \leq 1$. Table I lists normalized Stokes vectors for several polarization states.

The degree of polarization [Eq. (22)] can be written in terms of Stokes parameters as

$$P = \frac{\sqrt{s_1^2 + s_2^2 + s_3^2}}{s_0^2}. \tag{26}$$

Additionally, we can define the degree of linear polarization (the fraction of light in a linearly polarized state) by replacing the numerator of Eq. (26) with $\sqrt{s_1^2 + s_2^2}$, or the degree of circular polarization by replacing the numerator with s_3.

An optical component that changes the incident polarization state from $\mathbf{S}$ to some output state $\mathbf{S}'$ (through reflection, transmission, or scattering) can be described by a 4×4 Mueller matrix $\mathbf{M}$. This transformation is given by

$$\mathbf{S}' = \begin{bmatrix} s_0' \\ s_1' \\ s_2' \\ s_3' \end{bmatrix} = \mathbf{MS} = \begin{bmatrix} m_{00} & m_{01} & m_{02} & m_{03} \\ m_{10} & m_{11} & m_{12} & m_{13} \\ m_{20} & m_{21} & m_{22} & m_{23} \\ m_{30} & m_{31} & m_{32} & m_{33} \end{bmatrix} \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix}, \tag{27}$$

where $\mathbf{M}$ can be a product of n cascaded components $\mathbf{M}_i$ using

$$\mathbf{M} = \prod_{i=1}^{n} \mathbf{M}_i. \tag{28}$$

Matrix multiplication is not commutative and the product must be formed in the order that light reaches each component. For a system of three components in which the light is first incident on component 1 and ultimately exits component 3, $\mathbf{S}' = \mathbf{M}_3\mathbf{M}_2\mathbf{M}_1\mathbf{S}$, for example.

Examples of Mueller matrices for several homogeneous polarization components are given in Table II. The Mueller matrix for a component can be experimentally obtained by measuring $\mathbf{S}'$ for at least 16 judiciously selected $\mathbf{S}$ inputs, and procedures for measurement and data reduction are well developed.

C. Jones Calculus

In Jones calculus a two-element vector represents the amplitude and phase of the orthogonal electric field components and the phase information is preserved during calculation. This allows the coherent superposition of waves and is useful for describing the polarization state in systems such as interferometers that combine beams. Since this

TABLE I Matrix Representations of Selected Polarization States[a]

State	Coherency matrix	Stokes vector	Jones vector
Linear along $\hat{\mathbf{x}}$ ($\alpha = \beta = 0^\circ; \tan\epsilon = 0$)	$I\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$
Linear along $\hat{\mathbf{y}}$ ($\alpha = \beta = 90^\circ; \tan\epsilon = 0$)	$I\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
Linear at $+45^\circ$ ($\alpha = \beta = 45^\circ; \tan\epsilon = 0$)	$\frac{1}{2}I\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}$
General linear ($-90^\circ < \alpha < 90^\circ; \tan\epsilon = 0$)	$\frac{1}{2}I\begin{bmatrix} \cos(\alpha)^2 & \sin\alpha\cos\alpha \\ \sin\alpha\cos\alpha & \sin(\alpha)^2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ \cos 2\alpha \\ \sin 2\alpha \\ 0 \end{bmatrix}$	$\begin{bmatrix} \cos(\alpha) \\ \sin(\alpha) \end{bmatrix}$
Right circular ($\tan\epsilon = 1; \Delta\phi = 90^\circ; \beta = 45^\circ$)	$\frac{1}{2}I\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix}$
Left circular ($\tan\epsilon = -1; \Delta\phi = -90^\circ; \beta = 45^\circ$)	$\frac{1}{2}I\begin{bmatrix} 1 & \\ i & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -i \end{bmatrix}$
General elliptical		$\begin{bmatrix} 1 \\ \cos 2\epsilon\cos 2\beta \\ \cos 2\epsilon\sin 2\beta \\ \sin 2\epsilon \end{bmatrix}$	$\frac{1}{\sqrt{2}}\begin{bmatrix} \cos\beta e^{-i\Delta\phi/2} \\ \sin\beta e^{i\Delta\phi/2} \end{bmatrix}$
Unpolarized	$\frac{1}{2}I\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	None

[a] The parameters α, β, ϵ, and $\Delta\phi$ are defined corresponding to elliptical light as discussed in Section I. Extensive lists of Stokes and Jones vectors are available in several texts.

method is based on coherent waves, however, the Jones vector describes only fully polarized states, and partially or unpolarized states and depolarizing components cannot be represented.

Recalling Eqs. (1a) and (1b), one can write a vector formulation of complex representation for a fully coherent field

$$\mathbf{E} = e^{i\omega t}\begin{vmatrix} E_x e^{i\phi_x} \\ E_y e^{i\phi_y} \end{vmatrix}, \tag{29}$$

where the space-dependent term kz has been omitted. When the time dependence is also omitted, this vector is known as the full Jones vector. For generality, the Jones vector $\mathbf{J}$ is often written in a normalized form

$$\mathbf{J} = \begin{vmatrix} \cos\beta \\ \sin\beta e^{i\Delta\phi} \end{vmatrix} = \begin{vmatrix} \cos\beta e^{-i\Delta\phi/2} \\ \sin\beta e^{i\Delta\phi/2} \end{vmatrix}, \tag{30}$$

where $\Delta\phi = \phi_y - \phi_x$ and $\tan(\beta) = E_y/E_x$. The Jones vector can also be found from the polarization azimuthal angle

TABLE II Matrix Representation of Optical Components

Component	Mueller matrix	Jones matrix
Linear diattenuator with maximum (minimum) transmission p_1^2 (p_2^2) or absorber (if $p = p_1 = p_2$)	$\frac{1}{2}\begin{bmatrix} p_1+p_2 & p_1-p_2 & 0 & 0 \\ p_1-p_2 & p_1+p_2 & 0 & 0 \\ 0 & 0 & 2\sqrt{p_1 p_2} & 0 \\ 0 & 0 & 0 & 2\sqrt{p_1 p_2} \end{bmatrix}$	$\begin{bmatrix} p_1 & 0 \\ 0 & p_2 \end{bmatrix}$
Linear polarizer at 0°	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
Linear polarizer at an angle θ	$\frac{1}{2}\begin{bmatrix} 1 & \cos 2\theta & \sin 2\theta & 0 \\ \cos 2\theta & \cos^2 2\theta & \cos 2\theta \sin 2\theta & 0 \\ \sin 2\theta & \cos 2\theta \sin 2\theta & \sin^2 2\theta & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} \cos^2\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & \sin^2\theta \end{bmatrix}$
Half-wave ($\delta = 180°$) linear retarder with the fast axis at 0°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
Quarter-wave ($\delta = 90°$) linear retarder with the fast axis at 0°	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$	$\begin{bmatrix} e^{i\pi/4} & 0 \\ 0 & e^{-i\pi/4} \end{bmatrix}$
General linear retarder: retardance δ, fast axis at angle β from x axis	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 4\beta \sin^2\delta/2 + \cos^2\delta/2 & \sin 4\beta \sin^2\delta/2 & -\sin 2\beta \sin\delta \\ 0 & \sin 4\beta \sin^2\delta/2 & -\cos 4\beta \sin^2\delta/2 + \cos^2\delta/2 & \cos 2\beta \sin\delta \\ 0 & \sin 2\beta \sin\delta & -\cos 2\beta \sin\delta & \cos\delta \end{bmatrix}$	$\begin{bmatrix} e^{i\delta/2}\cos^2\beta + e^{-i\delta/2}\sin^2\beta & i\sin 2\beta \sin\delta/2 \\ i\sin 2\beta \sin\delta/2 & e^{-i\delta/2}\cos^2\beta + e^{i\delta/2}\sin^2\beta \end{bmatrix}$
Right circular retardance δ or rotator with $\theta = \delta/2$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\delta/2 & \sin\delta/2 & 0 \\ 0 & -\sin\delta/2 & \cos\delta/2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} \cos\delta/2 & \sin\delta/2 \\ -\sin\delta/2 & \cos\delta/2 \end{bmatrix}$
Mirror	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$
Faraday mirror	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$	$\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$
Depolarizer	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$	None

α and ellipticity $\tan(\varepsilon)$ of the polarization ellipse using

$$\Delta\phi = \tan^{-1}\left[\frac{\tan(2\epsilon)}{\sin(2\alpha)}\right]$$

$$\beta = \frac{1}{2}\cos^{-1}[\cos(2\epsilon)\cos(2\alpha)]. \qquad (31)$$

Table I provides examples of Jones vectors for several polarization states.

The polarization properties of optical components can be represented as 2×2 Jones matrices (Table II). The output polarization state is $\mathbf{J}' = \mathbf{MJ}$, where the Jones matrix $\mathbf{M}$ may be constructed from a cascade of components $\mathbf{M}_i$ using Eq. (28). In general the matrices are not commutative and require the same ordering as in Mueller calculus, with the rightmost matrix representing the first element the light is incident upon, and so on.

Jones used this calculus to establish three theorems that describe the minimum number of optical elements needed to describe a cascade of many elements at a given wavelength:

1. A system of any number of linear retarders and rotators (circular retarders) can be reduced to a system composed of only one retarder and one rotator.
2. A system of any number of partial polarizers and rotators can be reduced to a system composed of only one partial polarizer and one rotator.
3. A system of any number of retarders, partial polarizers, and rotators can be reduced to a system composed of only two retarders, one partial polarizer, and, at most, one rotator.

The Jones matrices in Table II assume forward propagation. In some cases, for example, with nonreciprocal components such as Faraday rotators, backward propagation must be explicitly described. Furthermore, since fields are used to represent polarization states, the phase shift arising from normal-incidence reflection may be important. For propagation in reciprocal media, the transformation from the forward Jones matrix to the backward case is given by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}_{\text{forward}} \rightarrow \begin{bmatrix} a & -c \\ -b & d \end{bmatrix}_{\text{backward}}. \tag{32}$$

For nonreciprocal behavior, such as the Faraday effect, the transformation is instead

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}_{\text{forward}} \rightarrow \begin{bmatrix} a & -b \\ -c & d \end{bmatrix}_{\text{backward}}. \tag{33}$$

When $\mathbf{M}$ is composed of a cascade of $\mathbf{M}_i$ that include both reciprocal and nonreciprocal polarization elements, each matrix must be transformed and a new combined matrix calculated. Upon reflection, the light is now backward propagating and the Jones matrix can be transformed to the forward-propagating form (for direct comparison with the input vector, for example) by changing the sign of the second element; in other words,

$$\mathbf{J}_{\text{forward}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{J}_{\text{backward}} \tag{34}$$

The calculi discussed above are applicable to problems when the polarization properties are lumped, that is, the system consists of simple components such as ideal waveplates, rotators, and polarizers, etc. Because the Jones (or Mueller) matrix from a cascade of matrices depends on the order of multiplication, an optical component with intermixed polarization properties cannot generally be represented by the simple multiplication matrices representing each individual property. For example, a component in which both linear retardance (represented by Jones matrix $\mathbf{M}_{\text{L}}$) and circular retardance ($\mathbf{M}_{\text{C}}$) are both distributed throughout the element is not properly represented by either $\mathbf{M}_{\text{L}}\mathbf{M}_{\text{C}}$ or $\mathbf{M}_{\text{C}}\mathbf{M}_{\text{L}}$.

A method known as the Jones N-matrix formulation can be used to find a single Jones matrix that properly describes the distribution of multiple polarization properties. The N-matrix represents the desired property over a vanishingly small optical path. The differential N-matrices for each desired property can be summed and the combined properties found by an integration along the optical path. Tables of N-matrices and algorithms for calculating corresponding Jones matrices can be found in several references.

Jones and Mueller matrices can be related to each other under certain conditions. Jones matrices differing only in absolute phase (in other words, a phase common to both orthogonal eigenpolarizations) can be transformed into a unique Mueller matrix that will have up to seven independent elements, though the phase information will be lost. Thus Mueller matrices for distributed polarization properties can be derived from Jones matrices calculated using N-matrices. Conversely, nondepolarizing Mueller matrices [which satisfy the condition $\text{Tr}(\mathbf{M}\mathbf{M}^{\text{T}}) = 4m_{00}$, where $\mathbf{M}^{\text{T}}$ is the transpose of $\mathbf{M}$] can be transformed into a Jones matrix.

D. Poincaré Sphere

The Poincaré sphere provides a visual method for representing polarization states and calculating the effects of polarizing components. Each state of polarization is represented by a unique point on the sphere defined by its azimuthal angle α, the ellipticity $\tan|\varepsilon|$, and the handedness. Orthogonal polarizations occupy points at opposite ends of a sphere diameter. Propagation through retarders is represented by a sphere rotation that translates the polarization state from an initial point to a final polarization.

Figure 10 shows a Poincaré sphere with several polarizations labeled. Point x represents linear polarization along the x axis and point y represents y-polarized light. Right-circular polarization ($\tan\epsilon = 1$) lies at the north pole, and all polarizations above the equator are right-elliptical. Similarly, the south pole represents left-circular polarization ($\tan\epsilon = -1$), and states below the equator are left-elliptically polarized. (In many texts the locations of the circular states are reversed; while a source of confusion, this change is valid so long as other conventions are observed.)

In Fig. 10, a general polarization state with azimuthal angle α and ellipticity angle ϵ is represented by the point p with longitude 2α and latitude 2ϵ. Linear polarizations have zero ellipticity ($\tan|\epsilon| = 0$) and are located along the

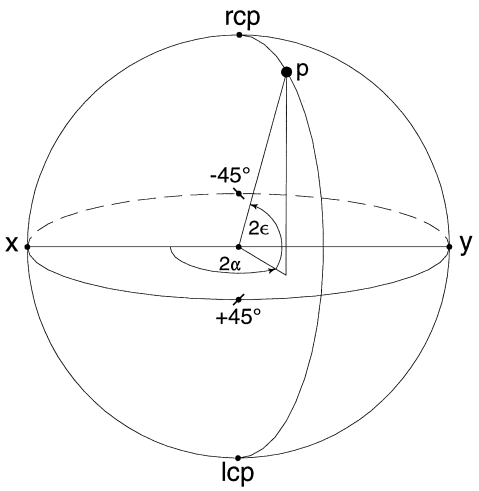


FIGURE 10 The Poincaré sphere. The polarization represented by point *p* is located using the azimuthal angle α (in the equatorial plane measured from point *x*) and the ellipticity angle ϵ (a meridional angle measured from the equator toward the north pole). Linear polarization along the *x* axis is located at point *x*, linear polarization along the *y* axis is represented by point *y*, and rcp and lcp denote right- and left-circularly polarized states, respectively. The origin represents unpolarized light.

equator. A linear polarization with azimuthal angle α from the x axis is located at a longitudinal angle 2α along the equator from point x. Polarization states that lie upon a circle parallel to the equator have the same ellipticity but different orientations. Polarizations at opposite diameters have the same ellipticity, perpendicular azimuthal angles, and opposite handedness.

The Poincaré sphere can also be used to show the effect of a retarder on an incident polarization state. A retarder oriented with a fast axis at α and an ellipticity and handedness given by $\tan\epsilon$ can be represented by a point R on the sphere located at angles 2α and 2ϵ. For a given input polarization represented by point p, a circle centered at point R that includes point p is the locus of the output polarization states possible for all retardance magnitudes. A specific retardance magnitude δ is represented by a clockwise arc of angle δ along the circle from the point p. The endpoint of this arc represents the polarization state output from the retarder.

Consider x-polarized light incident on a quarter-wave linear retarder oriented with its fast axis at $+45^\circ$ from horizontal; using Jones calculus, we find that right circular polarization should exit the waveplate. To show this graphically using the Poincaré sphere, we locate the point $+45^\circ$, which represents the retarder orientation. The initial polarization is at point x; for a retardance $\delta = 90^\circ$, we trace a clockwise arc centered at the point $+45^\circ$ that subtends 90° from point x. This arc ends at the north pole, so the resulting output is right-circular polarization. If the retardance was $\delta = 180^\circ$, the arc would subtend 180°, and the output light would be y-polarized. Similarly, left-circular polarization results if $\delta = 270^\circ$ (or if $\delta = 90^\circ$ and the fast axis is oriented at -45°). The evolution of the polarization through additional components can be traced by locating each retarder's representation on the sphere, defining a circle centered by this point and the polarization output from the previous retarder, and tracing a new arc through an angle equal to the retardance.

Comparing the Poincaré sphere definitions to Eq. (25) shows that for normalized Stokes vectors ($s_0 = 1$), each vector element corresponds to a point along Cartesian axes centered at the sphere's origin. Stokes element $s_1 (= \cos 2\epsilon \cos 2\alpha)$ falls along the axis between x- and y-polarized; $s_1 = 1$ corresponds to point x and $s_1 = -1$ corresponds to point y. Values of $s_2 (= \cos 2\epsilon \sin 2\alpha)$ correspond to points along the diameter connecting the $\pm 45^\circ$ linear polarization points; $s_2 = -1$ corresponds to the -45° point. Element $s_3 (= \sin 2\epsilon)$ is along the axis between the north and south poles. These projections on the Poincaré sphere can be equivalently represented by rewriting Eq. (25) and normalizing to obtain

$$\begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \cos(2\epsilon)\cos(2\alpha) \\ \cos(2\epsilon)\sin(2\alpha) \\ \sin(2\epsilon) \end{bmatrix}. \tag{35}$$

Any fully polarized state on the surface of the sphere can be found using these Cartesian coordinates. Partially polarized states will map to a point within the sphere, and unpolarized light is represented by the origin.

V. POLARIMETRY

Polarimetry is the measurement of a light wave's polarization state, or the characterization of an optical component's or material's polarization properties. Complete polarimeters measure the full Stokes vector of an optical beam or measure the full Mueller matrix of a sample. In many cases, however, some characteristics can be neglected and the measurement of all Stokes or Mueller elements is not necessary. Incomplete polarimeters measure a subset of characteristics and may be used when simplifying assumptions about the light wave (for example, that the degree of polarization is 1) or sample (for example, a retarder exhibits negligible diattenuation or depolarization) are appropriate. In this section, a few techniques are briefly described for illustration.

A. Light Measurement

A *polarization analyzer*, or light-measuring polarimeter, characterizes the polarization properties of an optical beam. An optical beam's Stokes vector can be completely

characterized by measuring the six optical powers listed in Eq. (25) using ideal polarizers. When the optical beam's properties are time invariant, the measurements can be performed sequentially by measuring the power transmitted through four orientations of a linear polarizer and two additional measurements with a quarter-wave retarder (oriented $\pm 45^\circ$ with respect to the polarizers axis) placed before the polarizer. In practice, as few as four measurements are required since $s_2 = 2P_{+45^\circ} - s_0$ and $s_3 = 2P_{\text{rcp}} - s_0$.

The Stokes vector can alternatively be measured with a single circular polarizer made by combining a quarter-wave plate (with the fast axis at 45°) with a linear polarizer. P_{rcp} is measured when the retarder side faces the source. Flipping so that the retarder faces the detector allows measurement of P_{0°, P_{90°, and $P_{\pm 45^\circ}$.

The Stokes vector elements can be measured simultaneously with multiple detector configurations. *Division of amplitude* polarimeters use beamsplitters to direct fractions of the power to appropriate polarization analyzers. Using *division of wavefront* polarization analyzers, we assume that the polarization is uniform over the optical beam and subdivisions of the beam's cross section are directed to appropriate analyzers.

Incomplete light-measuring polarimeters are useful when the light is fully polarized (degree of polarization approaches 1). For example, the ellipticity magnitude and azimuth can be found by analyzing the light with a rotating linear polarizer and measuring the minimum and maximum transmitted powers. Linear polarization yields a detected signal with maximum modulation, while minimum modulation occurs for circular polarization. The handedness of the ellipticity can be found using a right- (or left-) circular polarizer.

These methods are photometric, and accurate optical power measurements are required to determine the light characteristics. Before the availability of photodetectors, null methods that rely on adjusting system settings until light transmission is minimized were developed, and these are still useful today. For example, an incomplete polarimetric null system for analyzing polarized light uses a calibrated Babinet–Soleil compensator followed by a linear polarizer. Adjusting both the retardance δ and angle θ between the fast axis and polarizer axis until the transmitted power is zero yields the ellipticity angle ω (using $\sin 2\omega = \sin 2\theta \sin \delta$) and azimuthal angle α (using $\tan \alpha = \tan 2\theta \cos \delta$). When unpolarized light is present, the minimum transmission is not zero, and photometric measurement of this power can be used to obtain the degree of polarization.

B. Sample Measurement

A *polarization generator* is used to illuminate the sample with known states of polarization to measure the sample's polarization properties. The reflected or transmitted light is then characterized by a polarization analyzer, and the properties of the sample are inferred from changes between the input and output states.

A common configuration for determining the Mueller matrix combines a fixed linear polarizer and a rotating quarter-wave retarder for polarization generation with a rotating quarter-wave retarder followed by a fixed linear polarizer for analysis. Power is measured as the two retarders are rotated at different rates (one rotates five times faster than the other) and the Mueller matrix elements are found from Fourier analysis of the resulting time series. Alternatively, measurements can be taken at 16 (or more) specific combinations of generator and analyzer states, typically with the polarizers fixed and at specified retarder orientations. Data reduction techniques have been developed for efficiently determining the Mueller matrix from such measurements. Several methods include measurements at additional generator/analyzer combinations to overdetermine the matrix; least-squares techniques are then applied to reduce the influence of nonideal system components and decrease measurement error.

Because of the simplicity and reduction of variables, incomplete polarizers can often provide a more accurate measurement of a single polarization property when other characteristics are negligible. For example, there are many methods for measuring linear retardance in samples with negligible circular retardance, diattenuation, and depolarization, and these are often applicable to measurements of high-quality waveplates.

In a rotating analyzer system, the retarder is placed between two linear polarizers so that the input polarization bisects the retarder's birefringence axes. Linear retardance is calculated from measurements of the transmitted power when the analyzer is parallel (P_{0°) and perpendicular (P_{90°) to the input polarizer using $|\delta| = \cos^{-1}[(P_{0^\circ} - P_{90^\circ})/(P_{0^\circ} + P_{90^\circ})]$. In this measurement, retardance is limited to two quadrants (for example, measurements of 90° and $270^\circ = -90^\circ$ retarders will both yield $\delta = 90^\circ$). If a biasing quarter-wave retarder is placed between the input polarizer and retarder and both retarders are aligned with the fast axis at 45°, retardance in quadrants 1 and 4 ($|\delta| \leq 90^\circ$) can be measured from $\delta = \sin^{-1}[(P_{90^\circ} - P_{0^\circ})/(P_{90^\circ} + P_{0^\circ})]$. There are several null methods, including those that use a variable compensator aligned with the retarder at 45° between crossed polarizers (retardance is measured by adjusting a calibrated compensator until no light is detected) or that use a fixed quarter-wave-biasing retarder and rotate the polarizer and/or analyzer until a null is obtained.

Ellipsometry is a related technique that allows the measurement of isotropic optical properties of surfaces

and thin films from the polarization change induced upon reflection. Linearly polarized light is directed toward the sample at known incidence angles, and the reflected light is analyzed to determine its polarization ellipse.

Application of electromagnetic models to the configuration (for example, via Fresnel equations) allows one to calculate the refractive index, extinction coefficient, and film thickness from the measured ellipticities. Ellipsometry can be extended to other configurations using various incident polarizations and polarization analyzers to measure polarimetric quantities, blurring any distinction between ellipsometry and polarimetry.

SEE ALSO THE FOLLOWING ARTICLES

ELECTROMAGNETICS • LIGHT SOURCES • OPTICAL DIFFRACTION • WAVE PHENOMENA

BIBLIOGRAPHY

Anonymous (1984). "Polarization: Definitions and Nomenclature, Instrument Polarization," International Commission on Illumination, Paris.

Azzam, R. M. A., and Bashara, N. M. (1997). "Ellipsometry and Polarized Light," North Holland, Amsterdam.

Bennett, J. M. (1995). "Polarization." *In* "Handbook of Optics," Vol. 1, pp. 5.1–5.30 (Bass, M., ed.), McGraw-Hill, New York.

Bennett, J. M. (1995). "Polarizers." *In* "Handbook of Optics," Vol. 2, pp. 3.1–3.70 (Bass, M., ed.), McGraw-Hill, New York.

Born, M., and Wolf, E. (1980). "Principles of Optics," Pergamon Press, Oxford.

Chipman, R. A. (1995). "Polarimetry." *In* "Handbook of Optics," Vol. 1, pp. 22.1–22.37 (Bass, M., ed.), McGraw-Hill, New York.

Collet, E. (1993). "Polarized Light: Fundamentals and Applications," Marcel Dekker, New York.

Hecht, E., and Zajac, A. (1979). "Optics," Addison-Wesley, Reading, MA.

Kilger, D. S., Lewis, J. W., and Randall, C. E. (1990). "Polarized Light in Optics and Spectroscopy," Academic Press, San Diego, CA.

Yariv, A., and Yeh, P. (1984). "Optical Waves in Crystals," Wiley, New York.



Radiometry and Photometry

Ross McCluney
Florida Solar Energy Center

I. Background
II. Radiometry
III. Photometry
IV. Commonly Used Geometric Relationships
V. Principles of Flux Transfer
VI. Sources
VII. Optical Properties of Materials
VIII. The Detection of Radiation
IX. Radiometers and Photometers, Spectroradiometers, and Spectrophotometers
X. Calibration of Radiometers and Photometers

GLOSSARY

Illuminance, E_v The area density of luminous flux, the luminous flux per unit area at a specified point in a specified surface that is incident on, passing through, or emerging from that point in the surface (unit: $\mathrm{lm} \cdot \mathrm{m}^{-2} = \mathrm{lux}$).

Irradianc, E_e The area density of radiant flux, the radiant flux per unit area at a specified point in a specified surface that is incident on, passing through, or emerging from that point in the surface (unit: watt $\cdot\ \mathrm{m}^{-2}$).

Luminance, L_v The area and solid angle density of luminous flux, the luminous flux per unit projected area and per unit solid angle incident on, passing through, or emerging from a specified point in a specified surface, and in a specified direction in space (units: $\mathrm{lumen} \cdot \mathrm{m}^{-2} \cdot \mathrm{sr}^{-1} = \mathrm{cd} \cdot \mathrm{m}^{-2}$).

Luminous efficacy, K_r The ratio of luminous flux in lumens to radiant flux (total radiation) in watts in a beam of radiation (units: lumen/watt).

Luminous flux, Φ_v The $V(\lambda)$-weighted integral of the spectral flux Φ_λ over the visible spectrum (unit: lumen).

Luminous intensity, I_v The solid angle density of luminous flux, the luminous flux per unit solid angle incident on, passing through, or emerging from a point in space and propagating in a specified direction (units: $\mathrm{lm} \cdot \mathrm{sr}^{-1} = \mathrm{cd}$).

Photopic spectral luminous efficiency function, $V(\lambda)$ The standardized relative spectral response of a human observer under photopic (cone vision) conditions over the wavelength range of visible radiation.

Projected area, A_o Unidirectional projection of the area bounded by a closed curve in a plane onto another plane making some angle θ to the first plane.

Radiance, L_e The area and solid angle density of radiant flux, the radiant flux per unit projected area and per unit

solid angle incident on, passing through, or emerging from a specified point in a specified surface, and in a specified direction in space (units: watt $\cdot$ m$^{-2}\cdot$ sr^{-1}).

Radiant flux, Φ_e The time rate of flow of radiant energy (unit: watt).

Radiant intensity, I_e The solid angle density of radiant flux, the radiant flux per unit solid angle incident on, passing through, or emerging from a point in space and propagating in a specified direction (units: watt $\cdot$ sr^{-1}).

Solid angle, Ω The area A on a sphere of the radial projection of a closed curve in space onto that sphere, divided by the square r^2 of the radius of that sphere.

Spectral radiometric quantities The spectral "concentration" of quantity Q, denoted Q_λ, is the derivative $dQ/d\lambda$ of the quantity with respect to wavelength λ, where "Q" is any one of: radiant flux, irradiance, radiant intensity, or radiance.

RADIOMETRY is a system of language, mathematical formulations, and instrumental methodologies used to describe and measure the propagation of radiation through space and materials. The radiation so studied normally is confined to the ultraviolet (UV), visible (VIS), and infrared (IR) parts of the spectrum, but the principles are applicable to radiant energy of any form that propagates in space and interacts with matter in known ways, similar to those of electromagnetic radiation. This includes other parts of the electromagnetic spectrum and to radiation composed of the flow of particles where the trajectories of these particles follow known laws of ray optics, through space and through materials. Radiometric principles are applied to beams of radiation at a single wavelength or those composed of a broad range of wavelengths. They can also be applied to radiation diffusely scattered from a surface or volume of material. Application of these principles to radiation propagating through absorbing and scattering media generally leads to mathematically sophisticated and complex treatments when high precision is required. That important topic called *radiative transfer*, is not treated in this article.

Photometry is a subset of radiometry, and deals only with radiation in the visible portion of the spectrum. Photometric quantities are defined in such a way that they incorporate the variations in spectral sensitivity of the human eye over the visible spectrum, as a spectral weighting function built into their definition.

In determining spectrally broadband radiometric quantities, no spectral weighting function is used (or one may consider that a weighting "function" of unity (1.0) is applied at all wavelengths).

The scope of this treatment is limited to definitions of the primary quantities in radiometry and photometry, the derivations of several useful relationships between them, the rudiments of setting up problems in radiation transfer, short discussions of material properties in a radiometric context, and a very brief discussion of electronic detectors of electromagnetic radiation. The basic design of radiometers and photometers and the principles of their calibration are described as well.

Until the latter third of the 20th century, the fields of radiometry and photometry developed somewhat independently. Photometry was beset with a large variety of different quantities, names of those quantities, and units of measurement. In the 1960s and 1970s several authors contributed articles aimed at bringing order to the apparent confusion. Also, the International Lighting Commission (CIE, *Commission International de l'Eclairage*) and the International Electrotechnical Commission (CEI, *Commision Electrotechnic International*) worked to standardize a consistent set of symbols, units, and nomenclature, culminating in the *International Lighting Vocabulary*, jointly published by the CIE and the CEI. The recommendations of that publication are followed here. The CIE has become the primary international authority on terminology and basic concepts in radiometry and photometry.

I. BACKGROUND

A. Units and Nomenclature

Radiant flux is defined as the time rate of flow of energy through space. It is given the Greek symbol Φ and the metric unit watt (a joule of energy per second). An important characteristic of radiant flux is its distribution over the electromagnetic spectrum, called a *spectral distribution* or *spectrum*. The Greek symbol λ is used to symbolize the wavelength of monochromatic radiation, radiation having only one frequency and wavelength. The unit of wavelength is the meter, or a submultiple of the meter, according to the rules of *System International*, the international system of units (the metric system). The unit of frequency is the hertz (abbreviated Hz), defined to be a cycle (or period) per second. The symbol for frequency is the Greek ν. The relationship between frequency ν and wavelength λ is shown in the equation

$$\lambda\nu = c, \tag{1}$$

where c is the speed of propagation in the medium (called the "speed of light" more familarly). The spectral concentration of radiant flux at (or around) a given wavelength λ is given the symbol Φ_λ, the name *spectral radiant flux*, and the units watts per unit wavelength. An example of this is the watt per nanometer (abbreviated W/nm). The names, definitions, and units of additional radiometric quantities are provided in Section II.

The electromagnetic spectrum is diagramed in Fig. 1. The solar and visible spectral regions are expanded to the right of the scale. Though sound waves are not electromagnetic waves, the range of human-audible sound is shown in Fig. 1 for comparison.

The term "light" can only be applied in principle to electromagnetic radiation over the range of *visible* wavelengths. Radiation outside this range is invisible to the human eye and therefore cannot be called light. Infrared and ultraviolet radiation cannot be termed "light."

Names and spectral ranges have been standardized for the ultraviolet, visible, and infrared portions of the spectrum. These are shown in Table I.

TABLE I CIE Vocabulary for Spectral Regions

Name	Wavelength range
UV-C	100 to 280 nm
UV-B	280 to 315 nm
UV-A	315 to 400
VIS	Approx. 360–400 to 760–800 nm
IR-A[a]	780 to 1400 nm
IR-B	1.4 to 3.0 μm
IR-C[b]	3 μm to 1 mm

[a] Also called "near IR" or NIR.
[b] Also called "far IR" or FIR.

B. Symbols and Naming Conventions

When the wavelength symbol λ is used as a subscript on a radiometric quantity, the result denotes the concentration of the quantity at a specific wavelength, as if one were dealing with a monochromatic beam of radiation at this wavelength only. This means that the range $\Delta\lambda$ of wavelengths in the beam, around the wavelength λ of definition, is infinitesemally small, and can therefore be defined in terms of the mathematical derivative as follows. Let Q be a radiometric quantity, such as flux, and ΔQ be the amount of this quantity over a wavelength interval $\Delta\lambda$ centered at wavelength λ. The *spectral* version of quantity Q, at wavelength λ, is the derivative of Q with respect to wavelength, defined to be the limit as $\Delta\lambda$ goes to zero of the ratio $\Delta Q/\Delta\lambda$.

$$Q_\lambda = \frac{dQ}{d\lambda}. \tag{2}$$

This notation refers to the "concentration" of the radiometric quantity Q, at wavelength λ, rather than to its functional dependence on wavelength. The latter would be notated as $Q_\lambda(\lambda)$. Though seemingly redundant, this notation is correct within the naming convention established for the field of radiometry.

When dealing with the optical properties of *materials* rather than with concentrations of flux at a given wavelength, the subscripting convention is not used. Instead, the functional dependence on wavelength is notated directly, as with the spectral transmittance: $T(\lambda)$. *Spectral optical properties* such as this one are spectral weighting functions, not flux distributions, and their functional dependence on wavelength is shown in the conventional manner.

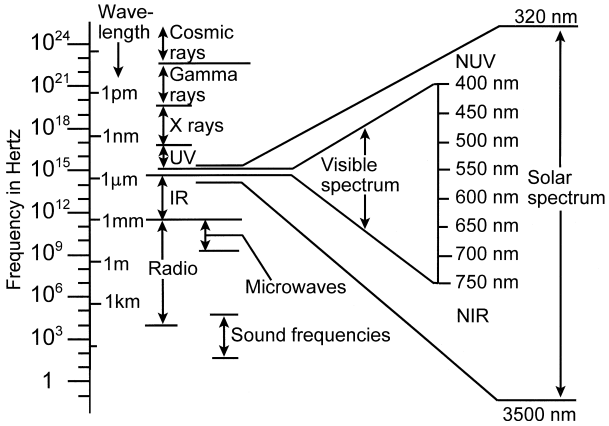


FIGURE 1 Wavelength and frequency ranges over the electromagnetic spectrum.

C. Geometric Concepts

In radiometry and photometry one is concerned with several geometrical constructs helpful in defining the spatial characteristics of radiation. The most useful are areas, plane angles, and solid angles.

The areas of interest are planar ones (including small differential elements of area used in definitions and derivations), nonplanar ones (areas on curved surfaces), and what are called *projected areas*. The latter are areas resulting when an original area is projected at some angle θ, as viewed from an infinite distance away. Projected areas are unidirectional projections of the area bounded by a closed curve in a plane onto another plane, one making angle θ to the first, as illustrated in Fig. 2.

A plane angle is defined by two straight lines intersecting at a point. The space between these lines in the plane defined by them is the plane angle. It is measured in radians (2π radians in a circle) or degrees (360 degrees to a circle). In preparation for defining solid angle it is pointed out that the plane angle can also be defined in terms of the radial projection of a line segment in a plane onto a point, as illustrated in Fig. 3.

> A *plane angle* is the quotient of the arc length s and the radius r of a radial projection of segment C of a curve in a plane onto a circle of radius r lying in that plane and centered at the vertex point P about which the angle is being defined.

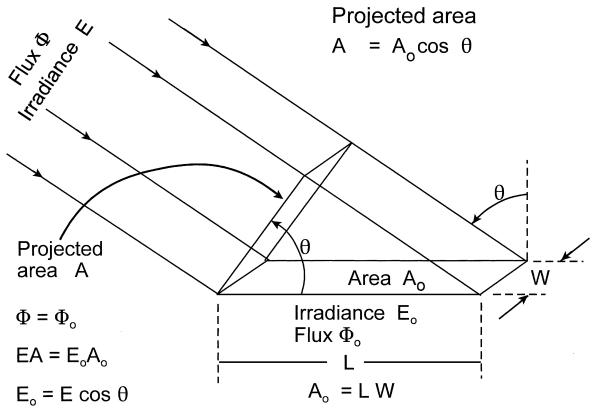


FIGURE 2 Illustration of the definition of projected areas.

If θ is the angle and s is the arc length of the projection onto a circle of radius r, then the defining equation is

$$\theta = \frac{s}{r}. \tag{3}$$

According to Eq. (3), the plane angle is a dimensionless quantity. However, to aid in communication, it has been given the unit *radian*, abbreviated *rad*. The radian measure of a plane angle can be converted to degree measure with the multiplication of a conversion constant, $180/\pi$.

A similar approach can be used to define *solid angle*. A solid angle is defined by a closed curve in space and a point, as illustrated in Fig. 4.

> A *solid angle* is the quotient of the area A and square of the radius r of a radial projection of a closed curve C in space onto a sphere of radius r centered at the vertex point P relative to which the angle is being defined.

If Ω is the solid angle being defined, A is the area on the sphere enclosed by the projection of the curve onto that sphere, and r is the sphere's radius, then the defining equation is

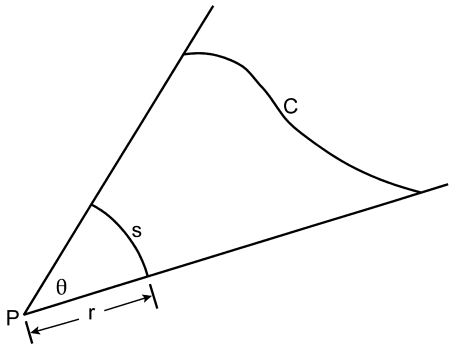


FIGURE 3 Definition of the plane angle.

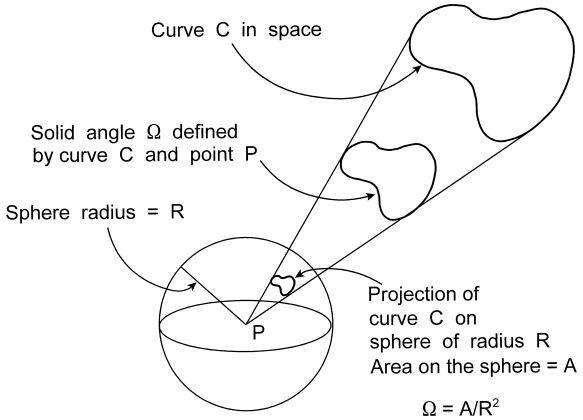


FIGURE 4 Definition of the solid angle.

$$\Omega = \frac{A}{r^2} \tag{4}$$

According to Eq. (4), the solid angle is dimensionless. However, to aid in communication, it has been given the unit *steradian*, abbreviated *sr*. Since the area of a sphere is 4π times the square of its radius, for a unit radius sphere the area is 4π and the solid angle subtended by it is 4π sr. The solid angle subtended by a hemisphere is $2\pi\,sr$. It is important to note that the area A in Eq. (4) is the area *on the sphere* of the projection of the curve C. It is not the area of a plane cut through the sphere and containing the projection of curve C. Indeed, the projections of some curves in space onto a sphere do not lie in a plane.

One which *does* is of particular interest—the projection of a circle in a plane perpendicular to a radius of the sphere, as illustrated in Fig. 5, which also shows a hemispherical solid angle. Let α be the plane angle subtended by the radius of the circle at the center of the sphere, called the "half-angle" of the cone. It can be shown that the solid angle Ω subtended by the circle is given by

$$\Omega = 2\pi(1 - \cos\alpha). \tag{5}$$

If $\alpha = 0$ then $\Omega = 0$ and if $\alpha = 90^\circ$ then $\Omega = 2\pi\,sr$, as required. A derivation of Eq. (5) is provided on pp. 28–30 of McCluney (1994).

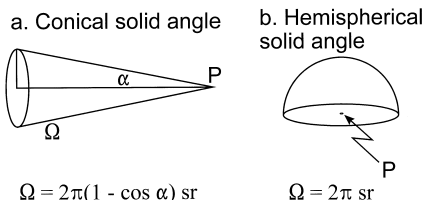


FIGURE 5 (a) Geometry for determining the solid angle of a right circular cone. α is the "half-angle" of the cone. (b) Geometry of a hemispherical solid angle.

D. The Metric System

To clarify the symbols, units, and nomenclature of radiometry and photometry the international system of units and related standards known as *the metric system* was embraced. There have been several versions of the metric system over the last couple of centuries. The current modernized one is named *Le System International d'Unites* (SI). It was established in 1960 by international agreement. The *Bureau International des Poids et Mesures* (BIPM) regularly publishes a document containing revisions and new recommendations on terminology and units. The International Standards Organization (ISO) publishes standards on the practical uses of the SI system in a variety of fields. Many national standards organizations around the world publish their own standards governing the use of this system, or translations of the BIPM documents, into the languages of their countries. In the United States the units metre and litre are spelled meter and liter, respectively.

The SI system calls for adherence to standard prefixes for standard orders of magnitude, listed in Table II. There are some simple rules governing the use of these prefixes. The prefix symbols are to be printed in roman type without spacing between the prefix symbol and the unit symbol. The grouped unit symbol plus its prefix is inseparable but may be raised to a positive or negative power and combined with other unit symbols. Examples: cm^2, nm, μm, klx, $1\ cm^2 = 10^{-4}\ m^2$. No more than one prefix can be used at a time. A prefix should never be used alone, except in descriptions of systems of units.

There are now two classes of units in the SI system:

- Base units and symbols: meter (m), kilogram (kg), second (s), ampere (A), kelvin (K), mole (mol), and candela (cd). Note that the abbreviations of units named for a person are capitalized, but the full unit name is not. (For example, the watt was named for James Watt and is abbreviated "W.")

TABLE II SI Prefixes

Factor	Prefix	Symbol	Factor	Prefix	Symbol
10^{24}	yotta	Y	10^{-1}	deci	d
10^{21}	zetta	Z	10^{-2}	centi	c
10^{18}	exa	E	10^{-3}	milli	m
10^{15}	peta	P	10^{-6}	micro	μ
10^{12}	tera	T	10^{-9}	nano	n
10^{9}	giga	G	10^{-12}	pico	p
10^{6}	mega	M	10^{-15}	femto	f
10^{3}	kilo	k	10^{-18}	atto	a
10^{2}	hecto	h	10^{-21}	zepto	z
10^{1}	decka, deca	da	10^{-24}	yocto	y

- Derived units: joule ($= kg \cdot m^2 \cdot s^{-2} = N \cdot m$), watt ($= J \cdot s^{-1}$), lumen ($= cd \cdot sr$), and lux ($= lm \cdot m^{-2}$). These are formed by combining base units according to algebraic relations linking the corresponding physical quantities. The laws of chemistry and physics are used to determine the algebraic combinations resulting in the derived units. Also included are the units of angle (radian, rad), and solid angle (steradian, sr).

A previously separate third class called supplementary units, combinations of the above units and units for plane and solid angle, was eliminated by the General Conference on Weights and Measures (CGPM, *Conference Generale des Poids et Mesures*) during its 9–12 October 1995 meeting. The radian and steradian were moved into the SI class of derived units.

Some derived units are given their own names, to avoid having to express every unit in terms of its base units. The symbol "·" is used to denote multiplication and "/" denotes division. Both are used to separate units in combinations. It is permissible to replace "·" with a space, but some standards require it to be included. In 1969 the following additional non-SI units were accepted by the International Committee for Weights and Measures for use with SI units: day, hour, and minute of time, degree, minute and second of angle, the litre ($10^{-3}\ m^3$), and the tonne (10^3 kg). In the United States the latter two are spelled "liter" and "metric ton," respectively.

The worldwide web of the internet contains many sites describing and explaining the SI system. A search on "The Metric System" with any search engine should yield several. The United States government site at http://physics.nist.gov/cuu/Units/ is comprehensive and provides links to other web pages of importance.

E. The I-P System

The most prominent alternative to the metric system is the inch-pound or the so-called "English" system of units. In this system the foot and pound are units for length and mass. The British thermal unit (Btu) is the unit of energy. This system is used little for radiometry and photometry around the world today, with the possible exception of the United States, where many illumination engineers still work with a mixed metric/IP unit, the foot-candle ($lumen \cdot ft^{-2}$) as their unit of illuminance. There are about 10.76 square feet in a square meter. So one foot-candle equals about 10.76 lux. The I-P system is being deprecated. However, in order to read older texts in radiometry and photometry using the I-P system, some familiarity with its units is advised. Tables 10.3 and 10.4 of McCluney (1994) provide conversion factors for many non-SI units.

II. RADIOMETRY

A. Definitions of Fundamental Quantities

There are five fundamental quantities of radiometry: radiant energy, radiant flux, radiant intensity, irradiance, and radiance. Each has a photometric counterpart, described in the next section.

Radiant energy, Q, is the quantity of energy propagating into, through, or emerging from a specified surface area in a specified period of time (unit: joule). Radiant energy is of interest in applications involving pulses of radiation, or exposure of a receiving surface to temporally continuous radiant energy over a specific period of time. An equivalent unit is the watt · sec.

Radiant flux (power), Φ, is the time rate of flow of radiant energy (unit: watt). One watt is 1 J sec^{-1}. The defining equation is the derivative of the radiant energy Q with respect to time t.

$$\Phi = \frac{dQ}{dt}. \tag{6}$$

Radiant flux is the quantity of energy passing through a surface or region of space per unit time. When specifying a radiant flux value, the spatial extent of the radiation field included in the specification should be described.

Irradiance, E, is the area density of radiant flux, the radiant flux per unit area at a specified point in a specified surface that is incident on, passing through, or emerging from that point in the surface (unit: watt $\cdot\, \text{m}^{-2}$). All directions in the hemispherical solid angle producing the radiation at that point are to be included. The defining equation is

$$E = \frac{d\Phi}{ds_o}, \tag{7}$$

where $d\Phi$ is an infinitesimal element of radiant flux and ds_o is an element of area in the surface. (The subscript "o" is used to indicate that this area is in an actual surface and is not a projected area.) The flux incident on a point in a surface can come from any direction in the hemispherical solid angle of incidence, or all of them, with any directional distribution. The flux can also be that leaving the surface in any direction in the hemispherical solid angle of emergence from the surface.

The irradiance leaving a surface can be called the *exitance* and can be given the symbol M, to distinguish it from the irradiance incident on the surface, but it has the same units and defining equation as irradiance. (The term *emittance*, related to the *emissivity*, is reserved for use in describing a dimensionless optical property of a material's surface and cannot be used for emitted irradiance.)

Since there is no mathematical or physical distinction between flux incident upon, passing through, or leaving a surface, the term irradiance is used throughout this article to describe the flux per unit area in all three cases.

Irradiance is a function of position in the surface specified for its definition.

When speaking of irradiance, one should be careful both to describe the surface and to indicate at which point on the surface the irradiance is being evaluated, unless this is very clear in the context of the discussion, or if the irradiance is known or assumed to be constant over the whole surface.

Radiant intensity, I, is the solid angle density of radiant flux, the radiant flux per unit solid angle incident on, passing through, or emerging from a point in space and propagating in a specified direction (units: watt $\cdot\, \text{sr}^{-1}$). The defining equation is

$$I = \frac{d\Phi}{d\omega}, \tag{8}$$

where $d\Phi$ is an element of flux incident on or emerging from a point within element $d\omega$ of solid angle in the specified direction. The representation of $d\omega$ in spherical coordinates is illustrated in Fig. 6.

Radiant intensity is a function of direction from its point of specification, and may be written as $I(\theta, \phi)$ to indicate its dependence upon the spherical coordinates (θ, ϕ) specifying a direction in space. Its definition is illustrated in Fig. 7.

Intensity is a useful concept for describing the directional distribution of radiation from a point source (or a source very small compared with the distance from it to the observer or detector of that radiation). The concept can be applied to extended sources having the same intensity at all points, in which case it refers to that subset of the radiation emanating from the entire source of finite and known area which flows into the same infinitesimal solid angle direction for each point in that area. (The next quantity

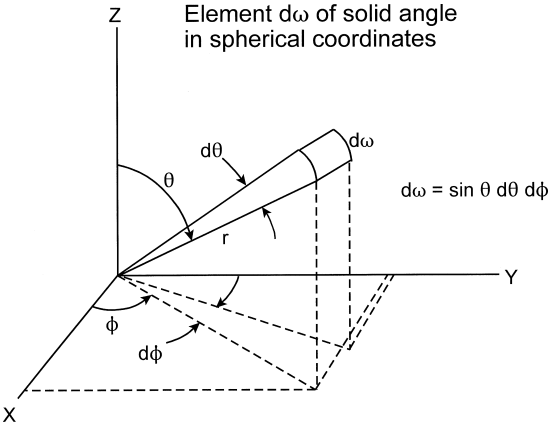


FIGURE 6 Representation of the element of solid angle $d\omega$ in Spherical coordinates.

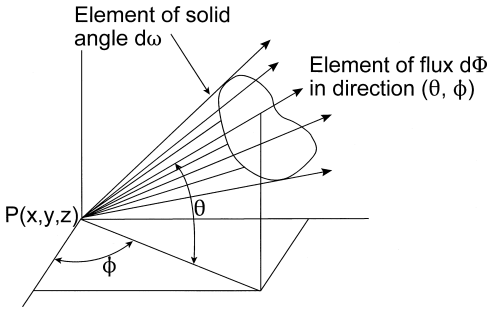


FIGURE 7 Geometry for the definition of Intensity.

to be described, *radiance*, is generally a more appropriate quantity for describing the directional distribution of radiation from nonpoint sources.)

When speaking of intensity, one should be careful to describe the point of definition and the direction of radiation from that point for clarity of discourse, unless this is obvious in the context of the discussion, or if it is known that the intensity is constant for all directions.

The word "intensity" is frequently used in optical physics. Most often the radiometric quantity being described is not intensity but irradiance.

Radiance, L, is the area and solid angle density of radiant flux, the radiant flux per unit projected area and per unit solid angle incident on, passing through, or emerging from a specified point in a specified surface, and in a specified direction (units: watt $\cdot$ m^{-2} $\cdot$ sr^{-1}). The defining equation is

$$L = \frac{d^2\Phi}{d\omega\, ds} \tag{9}$$

or

$$L = \frac{d^2\Phi}{d\omega\, ds_o \cos\theta},$$

where $ds = ds_o \cos\theta$ is the projected area, the area of the projection of elemental area ds_o along the direction of propagation to a plane perpendicular to this direction, $d\omega$ is an element of solid angle in the specified direction and θ is the angle this direction makes with the normal (perpendicular) to the surface at the point of definition, as illustrated in Fig. 8.

Radiance is a function of both position and direction. For many real sources, it is a strongly varying function of direction. It is the most general quantity for describing the propagation of radiation through space and transparent or semitransparent materials. The radiant flux, radiant intensity, and irradiance can be derived from the radiance by the mathematical process of integration over a finite surface area and/or over a finite solid angle, as demonstrated in Section IV.B.

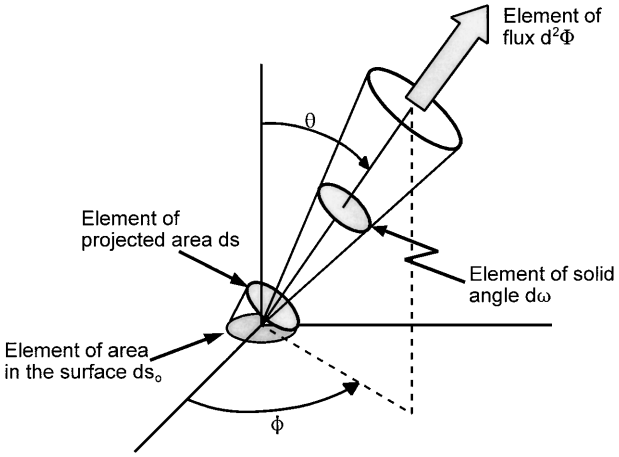


FIGURE 8 Geometry for the definition of radiance.

Since radiance is a function of position in a defined surface as well as direction from it, it is important when speaking of radiance to specify the surface, the point in it, and the direction from it. All three pieces of information are important for the proper specification of radiance. For example, we may wish to speak of the radiance emanating from a point on the ground and traveling upward toward the lens of a camera in an airplane or satellite traveling overhead. We specify the location of the point, the surface from which the flux emanates, and the direction of its travel toward the center of the lens. Since the words "radiance" and "irradiance" can sound very similar in rapidly spoken or slurred English, one can avoid confusion by speaking of the point and the surface that is common to both concepts, and then to clearly specify the direction when talking about radiance.

B. Definitions of Spectral Quantities

The spectral or wavelength composition of the five fundamental quantities of radiometry is often of interest. We speak of the "spectral distribution" of the quantities and by this is meant the possibly varying magnitudes of them at different wavelengths or frequencies over whatever spectral range is of interest. As before, if we let Q represent any one of the five radiometric quantities, we define the spectral "concentration" of that quantity, denoted Q_λ, to be the derivative of the quantity with respect to wavelength λ. (The derivative with respect to frequency ν or wavenumber $(1/\nu)$ is also possible but less used.)

$$Q_\lambda = \frac{dQ}{d\lambda}. \tag{10}$$

This defines the radiometric "quantity" per unit wavelength interval and can also be called the spectral power density. It has the same units as those of the quantity Q

TABLE III Symbols and Units of the Five Spectral Radiometric Quantities

Quantity	Spectral radiant energy	Spectral radiant flux	Spectral irradiance	Spectral intensity	Spectral radiance
Symbol	Q_λ	Φ_λ	E_λ	I_λ	L_λ
Units	$J \cdot nm^{-1}$	$W \cdot nm^{-1}$	$W \cdot m^{-2} \cdot nm^{-1}$	$W \cdot sr^{-1} \cdot nm^{-1}$	$W \cdot m^{-2} \cdot sr^{-1} \cdot nm^{-1}$

divided by wavelength. The spectral radiometric quantity Q_λ is in one respect the more fundamental of the two, since it contains more information, the spectral distribution of Q, rather than just its total magnitude. The two are related by the integral

$$Q = \int_0^\infty Q_\lambda \, d\lambda. \tag{11}$$

If Q_λ is zero outside some wavelength range, (λ_1, λ_2) then the integral of Eq. (11) can be replaced by

$$Q = \int_{\lambda_1}^{\lambda_2} Q_\lambda \, d\lambda. \tag{12}$$

The symbols and units of the spectral radiant quantities are listed in Table III.

III. PHOTOMETRY

A. Introduction

Photometry is a system of language, mathematical formulations, and instrumental methodologies used to describe and measure the propagation of *light* through space and materials. In consequence, the radiation so studied is confined to the visible (VIS) portion of the spectrum. Only light is visible radiation.

In photometry, all the radiant quantities defined in Section II are adapted or specialized to indicate the human eye's response to them. This response is built into the definitions. Familiarity with the five basic radiometric quantities introduced in that section makes much easier the study of the corresponding quantities in photometry, a subset of radiometry.

The human eye responds only to light having wavelengths between about 360 and 800 nm. Radiometry deals with electromagnetic radiation at all wavelengths and frequencies, while photometry deals only with visible light—that portion of the electromagnetic spectrum which stimulates vision in the human eye.

Radiation having wavelengths below 360 nm, down to about 100 nm, is called *ultraviolet*, or UV, meaning "beyond the violet." Radiation having wavelengths greater than 830 nm, up to about 1 mm, is called *infrared*, or IR, meaning "below the red." "Below" in this case refers to the *frequency* of the radiation, not to its wavelength. (Solving (1) for frequency yields the equation $\nu = c/\lambda$, showing the inverse relationship between frequency and wavelength.) The infrared portion of the spectrum lies *beyond* the red, having frequencies below and wavelengths above those of red light. Since the eye is very insensitive to light at wavelengths between 360 and about 410 nm and between about 720 and 830 nm, at the edges of the visible spectrum, many people cannot see radiation in portions of these ranges. Thus, the visible edges of the UV and IR spectra are as uncertain as the edges of the VIS spectrum.

The term "light" should only be applied to electromagnetic radiation in the visible portion of the spectrum, lying between 380 and 770 nm. With this terminology, there is no such thing as "ultraviolet light," nor does the term "infrared light" make any sense either. Radiation outside these wavelength limits is radiation—not light—and should not be referred to as light.

B. The Sensation of Vision

After passing through the cornea, the aqueous humor, the iris and lens, and the vitreous humor, light entering the eye is received by the retina, which contains two general classes of receptors: rods and cones. Photopigments in the outer segments of the rods and cones absorb radiation and the absorbed energy is converted within the receptors, into neural electrochemical signals which are then transmitted to subsequent neurons, the optic nerve, and the brain.

The cones are primarily responsible for day vision and the seeing of color. Cone vision is called *photopic* vision. The rods come into play mostly for night vision, when illumination levels entering the eye are very low. Rod vision is called *scotopic* vision. An individual's relative sensitivity to various wavelengths is strongly influenced by the absorption spectra of the photoreceptors, combined with the spectral transmittance of the preretinal optics of the eye. The relative spectral sensitivity depends on light level and this sensitivity shifts toward the blue (shorter wavelength) portion of the spectrum as the light level drops, due to the shift in spectral sensitivity when going from cones to rods.

The spectral response of a human observer under photopic (cone vision) conditions was standardized by the International Lighting Commission the *International de l'Eclairage* (CIE), in 1924. Although the actual spectral response of humans varies somewhat from person to person, an agreed standard response curve has been adopted,

as shown graphically in Fig. 9 and listed numerically in Table IV.

The values in Table IV are taken from the *Lighting Handbook* of the Illuminating Engineering Society of North America (IESNA). Since the symbol $V(\lambda)$ is normally used to represent this spectral response, the curve in Fig. 9 is often called the "V-lambda curve."

The 1924 CIE spectral luminous efficiency function for photopic vision defines what is called "the CIE 1924 Standard Photopic Photometric Observer." The official values were originally given for the wavelength range from 380 to 780 nm at 10-nm intervals but were then "completed by interpolation, extrapolation, and smoothing from earlier values adopted by the CIE in 1924 and 1931" to the wavelength range from 360 to 830 nm on 1-nm intervals and these were then recommended by the International Committee of Weights and Measures (CIPM) in 1976. The values below 380 and above 769 are so small to be of little value for most photometric calculations and are therefore not included in Table IV.

Any individual's eye may depart somewhat from the response shown in Fig. 9, and when light levels are moderately low, the other set of retinal receptors (rods) comes into use. This regime is called "scotopic vision" and is characterized by a different relative spectral response.

The relative spectral response curve for scotopic vision is similar in shape to the one shown in Fig. 9, but the peak is shifted from 555 to about 510 nm. The lower wavelength cutoff in sensitivity remains at about 380 nm, however, while the upper limit drops to about 640 nm. More information about scotopic vision can be found in various books on vision as well as in the *IESNA Lighting Handbook*. The latter contains both plotted and tabulated values for the scotopic spectral luminous efficiency function.

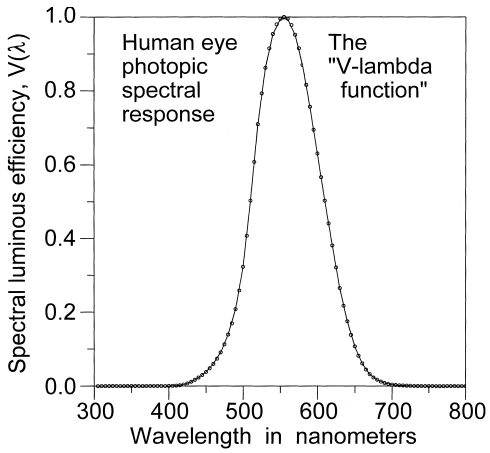


FIGURE 9 Human photopic spectral luminous efficiency.

C. Definitions of Fundamental Quantities

Five fundamental quantities in radiometry were defined in Section II.A. The photometric ones corresponding to the last four are easily defined in terms of their radiometric counterparts as follows. Let $Q_\lambda(\lambda)$ be one of the following: spectral radiant flux Φ_λ, spectral irradiance E_λ, spectral intensity I_λ, or spectral radiance L_λ. The corresponding photometric quantity, Q_v is defined as follows:

$$Q_\mathrm{v} = 683 \int_{380}^{770} Q_\lambda(\lambda) V(\lambda)\, d\lambda \tag{13}$$

with wavelength λ having the units of nanometers.

The subscript v (standing for "visible" or "visual") is placed on photometric quantities to distinguish them from radiometric quantities, which are given the subscript e (standing for "energy"). These subscripts may be dropped, as they were in previous sections, when the meaning is clear and no ambiguity results. Four fundamental radiometric quantities, and the corresponding photometric ones, are listed in Table V, along with the units for each.

To illustrate the use of (13), the conversion from spectral irradiance to illuminance is given by

$$E_\mathrm{v} = 683 \int_{380}^{770} E_\lambda(\lambda)\ V(\lambda)\ d\lambda. \tag{14}$$

The basic unit of luminous flux, the lumen, is like a "light-watt." It is the luminous equivalent of the radiant flux or power. Similarly, luminous intensity is the photometric equivalent of radiant intensity. It gives the luminous flux in lumens emanating from a point, per unit solid angle in a specified direction, and therefore has the units of lumens per steradian or lm/sr, given the name *candela*. This unit is one of the seven base units of the metric system. More information about the metric system as it relates to radiometry and photometry can be found in Chapter 10 of McCluney (1994).

Luminous intensity is a function of direction from its point of specification, and may be written as $I_v(\theta, \phi)$ to indicate its dependence upon the spherical coordinates (θ, ϕ) specifying a direction in space, illustrated in Fig. 6.

Illuminance is the photometric equivalent of irradiance and is like a "light-watt per unit area." Illuminance is a function of position (x, y) in the surface on which it is defined and may therefore be written as $E_v(x, y)$. Most light meters measure illuminance and are calibrated to read in lux. The lux is an equivalent term for the lumen per square meter and is abbreviated *lx*.

In the inch-pound (I-P) system of units, the unit for illuminance is the lumen per square foot, or lumen $\cdot$ ft^{-2}, which also has the odd name "foot-candle," abbreviated "fc," even though connection with candles and the candela

TABLE IV Photopic Spectral Luminous Efficiency $V(\lambda)$

Wavelength λ, nm	Standard values	Values interpolated at intervals of 1 nm								
		1	2	3	4	5	6	7	8	9
380	.00004	.000045	.000049	.000054	.000058	.000064	.000071	.000080	.000090	.000104
390	.00012	.000138	.000155	.000173	.000193	.000215	.000241	.000272	.000308	.000350
400	.0004	.00045	.00049	.00054	.00059	.00064	.00071	.00080	.00090	.00104
410	.0012	.00138	.00156	.00174	.00195	.00218	.00244	.00274	.00310	.00352
420	.0040	.00455	.00515	.00581	.00651	.00726	.00806	000889	.00976	.01066
430	.0116	.01257	.01358	.01463	.01571	.01684	.01800	.01920	.02043	.02170
440	.023	.0243	.0257	.0270	.0284	.0298	.0313	.0329	.0345	.0362
450	.038	.0399	.0418	.0438	.0459	.0480	.0502	.0525	.0549	.0574
460	.060	.0627	.0654	.0681	.0709	.0739	.0769	.0802	.0836	.0872
470	.091	.0950	.0992	.1035	.1080	.1126	.1175	.1225	.1278	.1333
480	.139	.1448	.1507	.1567	.1629	.1693	.1761	.1833	.1909	.1991
490	.208	.2173	.2270	.2371	.2476	.2586	.2701	.2823	.2951	.3087
500	.323	.3382	.3544	.3714	.3890	.4073	.4259	.4450	.4642	.4836
510	.503	.5229	.5436	.5648	.5865	.6082	.6299	.6511	.6717	.6914
520	.710	.7277	.7449	.7615	.7776	.7932	.8082	.8225	.8363	.8495
530	.862	.8739	.8851	.8956	.9056	.9149	.9238	.9320	.9398	.9471
540	.954	.9604	.9961	.9713	.9760	.9083	.9480	.9873	.9902	.9928
550	.995	.9969	.9983	.9994	1.0000	1.0002	1.0001	.9995	.9984	.9969
560	.995	.9926	.9898	.9865	.9828	.9786	.9741	.9691	.9638	.9581
570	.952	.9455	.9386	.9312	.9235	.9154	.9069	.8981	.8890	.8796
580	.870	.8600	.8496	.8388	.8277	.8163	.8046	.7928	.7809	.7690
590	.757	.7449	.7327	.7202	.7076	.6949	.6822	.6694	.6565	.6437
600	.631	.6182	.6054	.5926	.5797	.5668	.5539	.5410	.5282	.5156
610	.503	.4905	.4781	.4568	.4535	.4412	.4291	.4170	.4049	.3929
620	.381	.3690	.3575	.3449	.3329	.3210	.3092	.2977	.2864	.2755
630	.265	.2548	.2450	.2354	.2261	.2170	.2082	.1996	.1912	.1830
640	.175	.1672	.1596	.1523	.1452	.1382	.1316	.1251	.1188	.1128
650	.107	.1014	.0961	.0910	.0862	.0816	.0771	.0729	.0688	.0648
660	.061	.0574	.0539	.0506	.0475	.0446	.0418	.0391	.0366	.0343
670	.032	.0299	.0280	.0263	.0247	.0232	.0219	.0206	.0194	.0182
680	.017	.01585	.01477	.01376	.01281	.011,92	.01108	.01030	.00956	.00886
690	.0082	.00759	.00705	.00656	.00612	.00572	.00536	.00503	.00471	.00440
700	.0041	.00381	.00355	.00332	.00310	.00291	.00273	.00256	.00241	.00225
710	.0021	.001954	.001821	.001699	.001587	.001483	.001387	.001297	.001212	.001130
720	.00105	.000975	.000907	.000845	.000788	.000736	.000668	.000644	.000601	.000560
730	.00052	.000482	.000447	.000415	.000387	.000360	.000335	.000313	.000291	.000270
740	.00025	.000231	.000214	.000198	.000185	.000172	.000160	.000149	.000139	.000130
750	.00012	.000111	.000103	.000096	.000090	.000084	.000078	.000074	.000069	.000064
760	.00006	.000056	.000052	.000048	.000045	.000042	.000039	.000037	.000035	.000032

is mainly historical and indirect. The I-P system is being discontinued in photometry, to be replaced by the metric system, used exclusively in this treatment. For more information on the connections between modern metric photometry and the antiquated and deprecated units, the reader is directed to Chapter 10 of McCluney (1994). As with radiant exitance, illuminance leaving a surface can be called luminous exitance.

Luminance can be thought of as "photometric brightness," meaning that it comes relatively close to describing physically the subjective perception of "brightness." Luminance is the quantity of light flux passing through a point in a specified surface in a specified direction, per unit projected area at the point in the surface and per unit solid angle in the given direction. The units for luminance are therefore $\text{lm} \cdot \text{m}^{-2} \cdot \text{sr}^{-1}$. A more common unit for

TABLE V Basic Quantities of Radiometry and Photometry

Radiometric quantity	Symbol	Units	Photometric quantity	Symbol	Units
Radiant flux	Φ_e	watt (W)	Luminous flux	Φ_v	lumen (lm)
Radiant intensity	I_e	W/sr	Luminous intensity	I_v	lumen/sr = candela (cd)
Irradiance	E_e	W/m^2	Illuminance	E_v	$lumen/m^2 = lux$ (lx)
Radiance	L_e	$W \cdot m^{-2} \cdot sr^{-1}$	Luminance	L_v	$lm \cdot m^{-2} \cdot sr^{-1} = cd/m^2$

luminance is the $cd \cdot m^{-2}$, which is the same as the lumen per steradian and per square meter.

D. Luminous Efficacy of Radiation

Radiation luminous efficiacy, K_r, is the ratio of luminous flux (light) in lumens to radiant flux (total radiation) in watts in a beam of radiation. It is an important concept for converting between radiometric and photometric quantities. Its units are the lumen per watt, lm/W.

Luminous efficacy is not an efficiency since it is not a dimensionless ratio of energy input to energy output—it is a measure of the effectiveness of a beam of radiation in stimulating the perception of light in the human eye. If Q_v is any of the four photometric quantities (Φ_v, E_v, I_v, or L_v) defined previously and Q_e is the corresponding radiometric quantity, then the luminous efficacy associated with these quantities has the following defining equation:

$$K_r = \frac{Q_v}{Q_e} [lm \cdot W^{-1}] \tag{15}$$

Q_e is an integral over all wavelengths for which Q_λ is nonzero, while Q_v depends on an integral (13) over only the visible portion of the spectrum, where $V(\lambda)$ is nonzero. The luminous efficacy of a beam of infrared-only radiation is zero since none of the flux in the beam is in the visible portion of the spectrum. The same can be said of ultraviolet-only radiation.

The International Committee for Weights and Measures (CPIM), meeting at the International Bureau of Weights and Measures near Paris, France, in 1977 set the value 683 lm/W for the spectral luminous efficacy (K_r) of monochromatic radiation having a wavelength of 555 nm in standard air. In 1979 the candela was redefined to be the luminous intensity in a given direction, of a source emitting monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of $1/683$ W/sr. The candela is one of the seven fundamental units of the metric system. As a result of the redefinition of the candela, the value 683 shown in Eq. (13) is not a recommended good value for K_r but instead follows from the definition of the candela in SI units. (Prior to 1979, the candela was realized by a platinum approximation to a blackbody. After the 1979 redefinition of the candela, it can be realized from the absolute radiometric scale using any of a variety of absolute detection methods discussed in Section X.)

IV. COMMONLY USED GEOMETRIC RELATIONSHIPS

There are several important spatial integrals which can be developed from the definitions of the principal radiometric and photometric quantities. This discussion of some of them will use radiometric terminology, with the understanding that the same derivations and relationships apply to the corresponding photometric quantities.

A. Lambertian Sources and the Cosine Law

To simplify some derivations, an important property, approximately exhibited by some sources and surfaces, is useful. Any surface, real or imaginary, whose radiance is independent of direction is said to be a *Lambertian radiator*. The surface can be self-luminous, as in the case of a source, or it can be a reflecting or transmitting one. If the radiance emanating from it is independent of direction, this radiation is considered to be Lambertian. A Lambertian radiator can be thought of as a window onto an isotropic radiant flux field.

Finite Lambertian radiators obey Lambert's cosine law, which is that the flux in a given direction leaving an element of area in the surface varies as the cosine of the angle θ between that direction and the perpendicular to the surface element: $d\Phi(\theta) = d\Phi(0) \cos\theta$. This is because the projected area in the direction θ decreases with the cosine of that angle. In the limit, when $\theta = 90$ degrees, the flux drops to zero because the projected area is zero.

There is another version of the cosine law. It has to do not with the radiance *leaving* a surface but with how radiation from a uniform and collimated beam (a beam with all rays parallel to each other and equal in strength) incident on a plane surface is distributed over that surface as the angle of incidence changes.

This is illustrated as follows: A horizontal rectangle of length L and width W receives flux from a homogeneous beam of collimated radiation of irradiance E, making an angle θ with the normal (perpendicular) to the plane of

the rectangle, as shown in Fig. 2. If Φ is the flux over the projected area A, given by E times A, this same flux Φ_o will be falling on the larger horizontal area $A_o = L \cdot W$, producing horizontal irradiance $E_o = \Phi_o/A$. The flux is the same on the two areas ($\Phi = \Phi_o$). Equating them gives

$$EA = E_o A_o. \tag{16}$$

But $A = A_o \cos\ \theta$ so that

$$E_o = E \cos\theta. \tag{17}$$

This is another way of looking at the cosine law. Although it deals with the irradiance falling on a surface, if the surface is perfectly transparent, or even imaginary, it will also describe the irradiances (or exitances) emerging from the other side of the surface.

B. Flux Relationships

Radiance and irradiance are quite different quantities. Radiance describes the angular distribution of radiation while irradiance adds up all this angular distribution over a specified solid angle Ω and lumps it together. The fundamental relationship between them is embodied in the equation

$$E = \int_\Omega L(\theta, \phi) \cos\theta\, d\omega \tag{18}$$

for a point in the surface on which they are defined. In this and subsequent equations, the lower case $d\omega$ is used to identify an *element* of solid angle $d\omega$ and the upper case Ω to identify a finite solid angle.

If $\Omega = 0$ in Eq. (18), there is no solid angle and there can be no irradiance! When we speak of a collimated beam of some given irradiance, say E_o, we are talking about the irradiance contained in a beam of nearly parallel rays, but which necessarily have some small angular spread to them, filling a small but finite solid angle Ω, so that (18) can be nonzero. A perfectly collimated beam contains no irradiance, because there is no directional spread to its radiation—the solid angle is zero. Perfect collimation is a useful concept for theoretical discussions, however, and it is encountered frequently in optics. When speaking of collimation in experimental situations, what is usually meant is "quasi-collimation," nearly perfect collimation.

If the radiance $L(\theta, \phi)$ in Eq. (18) is constant over the range of integration (over the hemispherical solid angle), then it can be removed from the integral and the result is

$$E = \pi L. \tag{19}$$

This result is obtained from Eq. (18) by replacing $d\omega$ with its equivalence in spherical coordinates, $\sin\theta\, d\theta\, d\phi$, and integrating the result over the angular ranges of 0 to 2π for ϕ and 0 to $\pi/2$ for θ.

A constant radiance surface is called a Lambertian surface so that (19) applies only to such surfaces.

It is instructive to show how Eq. (18) can be derived from the definition of radiance. Eq. (9) is solved for $d^2\Phi$ and the result divided by ds_o. Since $d^2\Phi/ds_o = dE$ by Eq. (7), we have

$$dE = L \cos\theta\, d\omega. \tag{20}$$

Integrating (20) yields (18).

Similarly, one can replace the quotient $d^2\Phi/d\omega$ with the differential dI [from Eq. (8)] in Eq. (9) and solve for dI. Integrating the result over the source area S_o yields

$$I = \int_{S_o} L \cos\theta\ ds_o. \tag{21}$$

Intensity is normally applied only to point sources, or to sources whose area S_o is small compared with the distance to them. However, Eq. (21) is valid, even for large sources, though it is not often used this way.

Solving (8) for $d\Phi = I d\omega$, writing $d\omega$ as da_o/R^2, and dividing both sides by da_o yields the expression

$$E = \frac{I\ d\omega}{da_o} = \frac{I\ da_o}{da_o R^2} = \frac{I}{R^2} \tag{22}$$

for the irradiance E a distance R from a point source of intensity I, on a surface perpendicular to the line between the point source and the surface where E is measured. This is an explicit form for what is known as "the inverse square law" for the decrease in irradiance with distance from a point source. The inverse square law is a consequence of the definition of solid angle and the "filling" of that solid angle with flux emanating from a point source.

Next comes the conversion from radiance L to flux Φ. Let the dependence of the radiance on position in a surface on which it is defined be indicated by generalized coordinates (u, v) in the surface of interest. Let the directional dependence be denoted by (θ, ϕ), so that L may be written as a function L(u, v, θ, ϕ) of position and direction. Solve (9), the definition of radiance, for $d^2\Phi$. The result is

$$d^2\,\Phi = L \cos\theta\, ds_o\, d\omega. \tag{23}$$

Integrating (23) over both the area S_o of the surface and the solid angle Ω of interest yields

$$\Phi = \int_{S_o} \int_\Omega L(u, v, \theta, \phi) \cos\theta\ d\omega\ ds_o. \tag{24}$$

In spherical coordinates, $d\omega$ is given by $\sin\theta\ d\theta\ d\phi$. Letting the solid angle Ω over which (23) is integrated extend to the full 2π sr of the hemisphere, we have the total flux emitted by the surface in all directions.

$$\Phi = \int_{S_o} \int_0^{2\pi} \int_0^{\frac{\pi}{2}} L(u, v, \theta, \phi)\ \cos\theta\ \sin\theta\ d\theta\ d\phi\ ds_o. \tag{25}$$

V. PRINCIPLES OF FLUX TRANSFER

Only the geometrical aspects of flux transfer through a lossless and nonscattering medium are of interest in this section. The effects of absorption and scattering of radiation as it propagates through a transparent or semitransparent medium from a source to a receiver are outside the scope of this article. The effects of changes in the refractive index of the medium, however, are dealt with in Section V.E.

All uses of flux quantities in this section refer to both their radiant (subscript e) and luminous (subscript v) versions. The subscripts are left off for simplicity. When the terms radiance and irradiance are mentioned in this section, the discussion applies equally to luminance and illuminance, respectively.

A. Source/Receiver Geometry

The discussion begins with the drawing of Fig. 10 and the definition of radiance L in (9):

$$L = \frac{d^2\Phi}{d\omega\, ds_o \cos\theta}, \tag{26}$$

where θ is the angle made by the direction of emerging flux with respect to the normal to the surface of the source, ds_o is an infinitesimally small element of area at the point of definition in the source, and $d\omega$ is an element of solid angle from the point of definition in the direction of interest.

In Fig. 10 are shown an infinitesimally small element ds_o of area at a point in a source, an infinitesimal element da_o of area at point P on a receiving surface, the distance R between these points, and the angles θ and ψ between the line of length R between the points and the normals to the surfaces at the points of intersection, respectively.

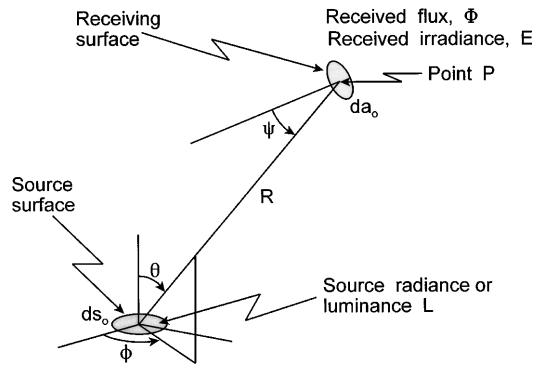


FIGURE 10 Source/receiver geometry.

B. Fundamental Equations of Flux Transfer

The element $d\omega$ of solid angle subtended by element of projected receiver area $da = da_o \cos\psi$ at distance R from the source is

$$d\omega = \frac{da}{R^2} = \frac{da_o \cos\psi}{R^2} \tag{27}$$

so that, solving (26) for $d^2\Phi$ and using (27), the element of flux received at point P from the element ds_o of area of the source is given by

$$d^2\Phi = L\frac{ds_o \cos\theta\; da_o \cos\psi}{R^2} \tag{28}$$

with the total flux Φ received by area A_o from source area S_o being given by

$$\Phi = \int_{S_o}\int_{A_o} L\frac{ds_o \cos\theta\; da_o \cos\psi}{R^2}. \tag{29}$$

This is the fundamental (and very general within the assumptions of this section) equation describing the transfer of radiation from a source surface of finite area to a receiving surface of finite area. Most problems of flux transfer involve this integration (or a related version shown later, giving the irradiance E instead of the flux). For complex or difficult geometries the problem can be quite complex analytically because in such cases L, θ, ψ, and R will be possibly complicated functions of position in both the source and the receiver surfaces. The general dependency of L on direction is also embodied in this equation, since the direction from a point in the source to a point in the receiver generally changes as the point in the receiver moves over the receiving surface.

The evaluation of (29) involves setting up diagrams of the geometry and using them to determine trigonometric and other analytic relationships between the geometric variables in (29). If the problem is expressed in cartesian coordinates, for example, then the dependences of L, θ, ψ, R, ds_o, and da_o upon those coordinates must be determined so that the integrals in (29) can be evaluated.

Two important simplifications allow us to address a large class of problems in radiometry and photometry with ease, by simplifying the mathematical analysis.

The first results when the source of radiance is known to be Lambertian and to have the same value for all points in the source surface. This makes L constant over all ranges of integration, both the integration over the source area and the one over the solid angle of emerging directions from each point on the surface. In such a case, the radiance can be removed from all integrals over these variables. The remaining integrals are seen to be purely geometric in character. The second simplification arises when one doesn't want the total flux over the whole receiving surface—only the flux per unit area at a point on that surface, the

irradiance E at point P in Fig. 10. In this case, we can divide both sides of (28) by the element of area in the receiving surface, da_o, to get

$$dE = L\frac{\cos\theta\cos\psi}{R^2}ds_o. \tag{30}$$

This equation is the counterpart of (28) when it is the irradiance E of the receiving surface that is desired. For the total irradiance at point P, one must integrate this equation over the portion S_o of the source surface contributing to the flux at P.

$$E = \int_{S_o} L\frac{\cos\theta\cos\psi}{R^2}ds_o. \tag{31}$$

When L is constant over direction, it can be removed from this integral and one is left with a simpler integration to perform. Equation (31) is the counterpart to (29) when it is the irradiance E at the receiving point that is of interest rather than the total flux over area A_o.

C. Simplified Source/Receiver Geometries

If the source area S_o is small with respect to the distance R to the point P of interest (i.e., if the maximum dimension of the source is small compared with R), then R^2, $\cos\psi$, and $\cos\theta$ do not vary much over the range of integration shown in (31) and they can be removed from the integral. If L does not vary over S_o, then it also can be removed from the integral, even if L is direction dependent, because the range of integration over direction is so small; that is, only the one direction from the source to point P in the receiver is of interest. We are left with an approximate version of (30) for small homogeneous sources some distance from the point of reception:

$$E \approx L\frac{S_o\cos\theta\cos\psi}{R^2}. \tag{32}$$

This equation contains within it both the cosine law and the inverse square law.

If the source and receiving surfaces face each other directly, so that θ and ψ are zero, both of the cosines in this equation have values of unity and the equation is still simpler in form.

D. Configuration Factor

In analyzing complicated radiation transfer problems, it is frequently helpful to introduce what is called the *configuration factor*. Alternate names for this factor are the *view*, *angle*, *shape*, *interchange*, or *exchange factor*. It is defined to be the fraction of total flux from the source surface that is received by the receiving surface. It is given the symbol F_{s-r} or F_{1-2}, indicating flux transfer from source to receiver or from Surface 1 to Surface 2. In essence, it indicates the details of how flux is transferred from a source area of some known form to a reception area. Its value is most evident when the source radiance is of such a nature that it can be taken from the integrals, leaving integrals over only geometric variables.

The geometry can still be quite complex, making analytical expressions for F_{1-2} difficult to determine and calculate. Many important geometries have already been analyzed, however, and the resulting configuration factors published.

In many problems, one is most concerned with the magnitude and spectral distribution of the source radiance and the corresponding spectral irradiance in a receiving surface, rather than with the geometrical aspects of the problem expressed by the shape factor. It is very convenient in such cases to separate the spectral variations from the geometrical ones. Once the configuration factor has been determined for a situation with nonchanging geometry, it remains constant and attention can be focused on the variable portion of the solution.

A general expression for the configuration factor results from dividing (29) for the flux Φ_r on the receiver by (24) for the total flux Φ_s, emitted by the source.

$$F_{s-r} = \frac{\Phi_r}{\Phi_s} = \frac{\int_{S_o}\int_{A_o}\dfrac{L\cos\theta\cos\psi\,ds_o\,da_o}{R^2}}{\int_{S_o}\int_{2\pi}L\cos\theta\,d\omega\,ds_o}. \tag{33}$$

This is the most general expression for the configuration factor. If the source is Lambertian and homogeneous, or if S_o and A_o are small in relation to R^2 then L can be removed from the integrals, resulting in

$$F_{s-r} = \frac{\int_{S_o}\int_{A_o}\dfrac{\cos\theta\cos\psi\,ds_o\,da_o}{R^2}}{\pi S_o} \tag{34}$$

a more conventional form for the configuration factor. As desired, it is purely geometric and has no radiation components. For homogeneous Lambertian sources of radiance L, the flux to a receiver, Φ_{s-r} is given by

$$\Phi_{s-r} = \pi\, S_o L F_{s-r} \tag{35}$$

E. Effect of Refractive Index Changes

For a ray propagating through an otherwise homogeneous medium without losses, it can be shown that the quantity L/n^2 is invariant along the ray. L is the radiance and n is the refractive index of the medium. If the refractive index is constant, the radiance L is constant along the ray. This is known as the *invariance of radiance*.

Suppose this ray passes through a (specular) interface between two isotropic and homogeneous media of different refractive indices, n_1 and n_2, and suppose there is neither absorption nor reflection at the interface. In this case it can be shown that

$$\frac{L_1}{n_1^2} = \frac{L_2}{n_2^2}. \tag{36}$$

Equation 36 shows how radiance invariance is modified for rays passing through interfaces between two media with different refractive indices.

A consequence of (36) is that a ray entering a medium of different refractive index will have its radiance altered, but upon emerging back into the original medium the original radiance will be restored, neglecting absorption, scattering, and reflection losses.

This is what happens to rays passing through the lens of an imaging system. The radiance associated with every ray contributing to a point in an image is the same as when that ray left the object on the other side of the lens (ignoring reflection and transmission losses in the lens). Since this is true of all rays making up an image point, the radiance of an image formed by a perfect, lossless lens equals the radiance of the object (the source).

This may seem paradoxical. Consider the case of a focusing lens, one producing a greater irradiance in the image than in the object. How can a much brighter image have the same radiance as that of the object? The answer is that the increased flux *per unit area* in the image is balanced by an equal reduction in the flux *per unit solid angle* incident on the image. This trading of flux per unit area for flux per unit solid angle is what allows the radiance to remain essentially unchanged.

VI. SOURCES

A. Introduction

The starting point in solving most problems of radiation transfer is determining the magnitude and the angular and spectral distributions of emission from the source. The optical properties of any materials on which that radiation is incident are also important, especially their spectral and directional properties. This section provides comparative information about a variety of sources commonly found in radiometric and photometric problems within the UV, VIS, and IR parts of the spectrum. Definitions used in radiometry and photometry for the reflection, transmission, and absorption properties of materials are provided in Section VII.

Spectral distributions are probably the most important characteristics of sources that must be considered in the design of radiometric systems intended to measure all or portions of those distributions. The matching of a proper detector/filter combination to a given radiation source is one of the most important tasks facing the designer. Section VIII deals with detectors.

B. Blackbody Radiation

All material objects above a temperature of absolute zero emit radiation. The hotter they are, the more they emit. The constant agitation of the atoms and molecules making up all objects involves accelerated motion of electrical charges (the electrons and protons of the constituent atoms). The fundamental laws of electricity and magnetism, as embodied in Maxwell's equations, predict that any accelerated motion of charges will produce radiation. The constant jostling of atoms and molecules in material substances above a temperature of absolute zero produces electromagnetic radiation over a broad range of wavelengths and frequencies.

1. Stefan–Boltzmann Law

The total radiant flux emitted from the surface of an object at temperature T is expressed by the Stefan–Boltzmann law, in the form

$$M_{\text{bb}} = \sigma T^4, \tag{37}$$

where M_{bb} is the exitance of (irradiance leaving) the surface in a vacuum, σ is the Stefan–Boltzmann constant ($5.67031 \times 10^{-8}\ \text{W} \cdot \text{m}^{-2} \cdot \text{K}^{-4}$), and T is the temperature in degrees kelvin. The units for M_{bb} in (37) are $\text{W} \cdot \text{m}^{-2}$. Using (37), a blackbody at 27°C, ($27 + 273 = 300$ K), emits at the rate of 460 W/m^2. At 100°C this rate increases to 1097 W/m^2.

Equation (37) applies to what is called a *perfect* or *full* emitter, one emitting the maximum quantity of radiation possible for a surface at temperature T. Such an emitter is called a *blackbody*, and its emitted radiation is called *blackbody radiation*.

A blackbody is defined as an ideal body that allows all incident radiation to pass into it (zero reflectance) and that absorbs internally all the incident radiation (zero transmittance). This must be true for all wavelengths and all angles of incidence. According to this definition, a blackbody is a perfect absorber, having absorptance 1.0 at all wavelengths and directions. Due to the law of the conservation of energy, the sum of the reflectance R and absorptance A of an opaque surface must be unity, $A + R = 1.0$. Thus, if a blackbody has an absorptance of 1.0, its reflectance must be zero. Accordingly, a perfect blackbody at room temperature would appear totally black to the eye, hence the origin of the name. Only a few surfaces, such as carbon black, carborundum, and gold black, approach a blackbody in these optical properties.

The radiation emitted by a surface is in general distributed over a range of angles filling the hemisphere and over a range of wavelengths. The angular distribution of radiance from a blackbody is constant; that is, the radiance is independent of direction; it is Lambertian. Specifically, this means that $L_\lambda(\theta, \phi) = L_\lambda(0, 0) = L_\lambda$. Thus, the relationship between the spectral radiance $L_{bb\lambda}$ and spectral exitance $M_{bb\lambda}$ of a blackbody is given by (19), repeated here as

$$M_{bb\lambda} = \pi L_{bb\lambda}. \tag{38}$$

If $L_{bb\lambda}$ is in $W \cdot m^{-2} \cdot sr^{-1} \cdot nm^{-1}$ then the units of $M_{bb\lambda}$ will be $W \cdot m^{-2} \cdot nm^{-1}$.

2. Greybodies

Imperfect emitters, which emit less than a blackbody at any given temperature, can be called *greybodies* if their spectral shape matches that of a blackbody. If that shape differs from that of a blackbody the emitter is called a *nonblackbody*.

The Stefan–Boltzmann law still applies to greybodies, but an optical property factor must be included in (37) and (38) for them to be correct for greybodies. That is the *emissivity* of the surface, defined and discussed in Section VII.C.3.

3. Planck's Law

As the temperature changes, the spectral distribution of the radiation emitted by a blackbody shifts. In 1901, Max Planck made a radical new assumption—that radiant energy is quantized—and used it to derive an equation for the spectral radiant energy density in a cavity at thermal equilibrium (a good theoretical approximation of a blackbody). By assuming a small opening in the side of the cavity and examining the spectral distribution of the emerging radiation, he derived an equation for the spectrum emitted by a blackbody. The equation, now called Planck's blackbody spectral radiation law, accurately predicts the spectral radiance of blackbodies in a vacuum at any temperature. Using the notation of this text the equation is

$$L_{bb\lambda} = \frac{2hc^2}{\lambda^5 \left(e^{\frac{hc}{\lambda kT}} - 1\right)}, \tag{39}$$

where $h = 6.626176 \times 10^{-34}$ J·s is Planck's constant, $c = 2.9979246 \times 10^8$ m/s is the speed of light in a vacuum, and $k = 1.380662 \times 10^{-23}\,J \cdot K^{-1}$ is Boltzmann's constant. Using these values, the units of $L_{bb\lambda}$ will be $W \cdot m^{-2} \cdot \mu m^{-1} \cdot sr^{-1}$. Plots of the spectral distribution of a blackbody for different temperatures are illustrated in Fig. 11. Each curve is labeled with its temperature in degrees Kelvin. Insignificant quantities of blackbody radiation lie in the visible portion of the spectrum for temperatures below about 1000 K. With increasing temperatures, blackbody radiation first appears red, then white, and at very high temperatures it has a bluish appearance.

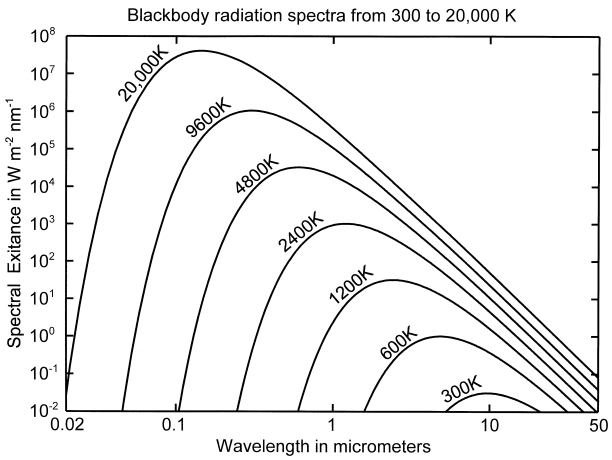


FIGURE 11 Exitance spectra for blackbodies at various temperatures from 300 to 20,000 K, calculated using Eq. (47).

From (38), the spectral exitance $M_{bb\lambda}$ of a blackbody at temperature T is just the spectral radiance $L_{bb\lambda}$ given in (39) multiplied by π.

$$M_{bb\lambda} = \frac{2\pi hc^2}{\lambda^5 \left(e^{\frac{hc}{\lambda kT}} - 1\right)}. \tag{40}$$

4. Luminous Efficacy of Blackbody Radiation

Substituting (40) for the hemispherical spectral exitance of a blackbody into (11) for E_e and (14) for E_v, for each of several different temperatures T, one can calculate the radiation luminous efficacy K_{bb} of blackbody radiation as a function of temperature. Some numerical results are given in Table VI, where it can be seen that, as expected, the luminous efficacy increases as the body heats up to white hot temperatures. At very high temperatures K_{bb} declines, since the radiation is then strongest in the ultraviolet, outside of the visible portion of the spectrum.

5. Experimental Approximation of a Blackbody

The angular and spectral characteristics of a blackbody can be approximated with an arrangement similar to the one shown in Fig. 12. A metal cylinder is hollowed out to form a cavity with a small opening in one end. At the opposite end is placed a conically shaped "light trap," whose purpose is to multiply reflect incoming rays, with maximum absorption at each reflection, in such a manner that a very large number of reflections must take place before any incident ray can emerge back out the opening. With the absorption high on each reflection, a vanishingly small

TABLE VI Blackbody Luminous Efficacy Values

Temperature in degrees K	Luminous efficacy in lm/W
500	7.6×10^{-13}
1,000	2.0×10^{-4}
1,500	0.103
2,000	1.83
2,500	8.71
3,000	21.97
4,000	56.125
5,000	81.75
6,000	92.9
7,000	92.8
8,000	87.3
9,000	79.2
10,000	70.6
15,000	37.1
20,000	20.4
30,000	7.8
40,000	3.7
50,000	2.0

fraction of incident flux, after being multiply reflected and scattered, emerges from the opening. In consequence, only a very tiny portion of the radiation passing into the cavity through the opening is reflected back out of the cavity.

The temperature of the entire cavity is controlled by heating elements and thick outside insulation so that all surfaces of the interior are at precisely the same (known) temperature and any radiation escaping from the cavity will be that emitted from the surfaces within the cavity. The emerging radiation will be rendered very nearly isotropic by the multiple reflections taking place inside (at least over the useful solid angle indicated in Fig. 12, for which the apparatus is designed).

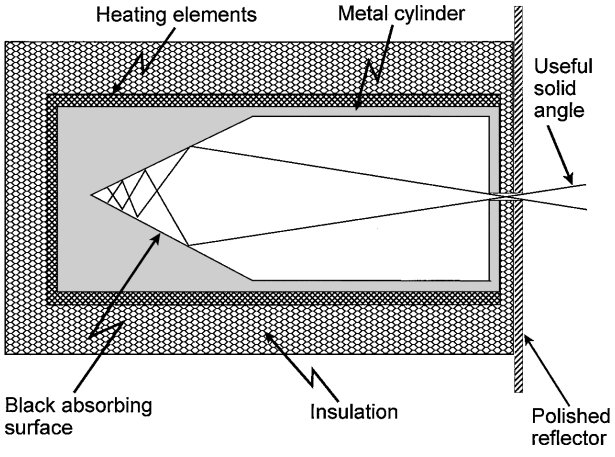


FIGURE 12 Schematic diagram of an approximation to a blackbody.

C. Electrically Powered Sources

Modern tungsten halogen lamps in quartz envelopes produce output spectra that are somewhat similar in shape to those of blackbody distributions. A representative spectral distribution is shown in Fig. 13. This lamp covers a wide spectral range, including the near UV, the visible, and much of the infrared portion of the spectrum. Only the region from about 240 to 2500 nm is shown in Fig. 13.

Although quartz halogen lamps produce usable outputs in the ultraviolet region, at least down to 200 nm, the output at these short wavelengths is quite low and declines rapidly with decreasing wavelength. Deuterium arc lamps overcome the limitations of quartz halogen lamps in this spectral region, and they do so with little output above a wavelength of 500 nm except for a strong but narrow emission line at about 660 nm. The spectral irradiance from a deuterium lamp is plotted in Fig. 14.

Xenon arc lamps have a more balanced output over the visible but exhibit strong spectral "spikes" that pose problems in some applications. Short arc lamps, such as those using xenon gas, are the brightest manufactured sources, with the exception of lasers. Because of the nature of the arc discharges, these lamps emit a continuum of output over wavelengths covering the ultraviolet and visible portions of the spectrum.

The spectral irradiance outputs of the three sources just mentioned cover the near UV, the visible, and the near IR. They are plotted along with the spectrum of a 50-W mercury-vapor arc lamp in Fig. 14. Mercury lamps emit strong UV and visible radiation, with strong spectral

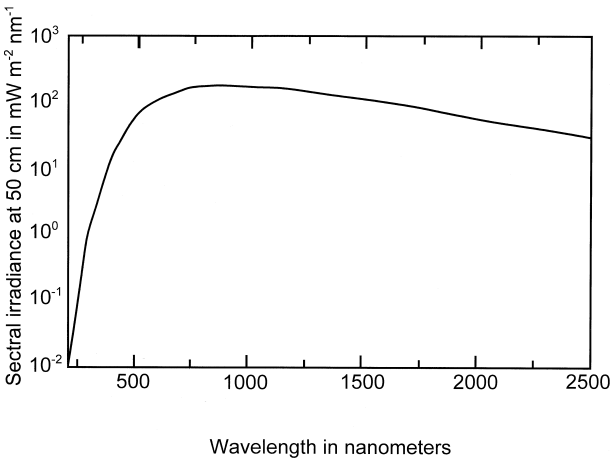


FIGURE 13 Spectral irradiance from a quartz halogen lamp 50 cm from the filament.

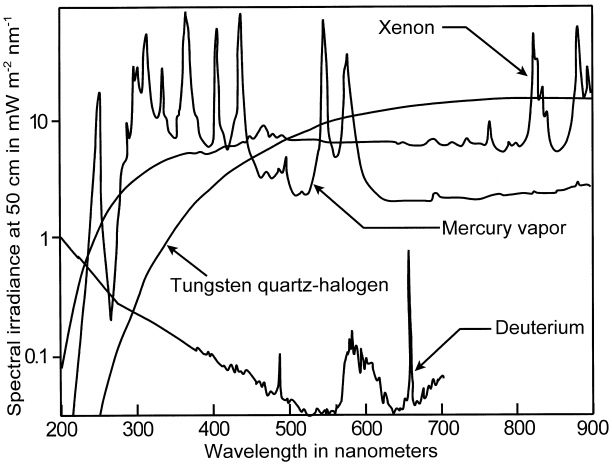


FIGURE 14 Spectral irradiance distributions for four sources of radiation.

lines in the ultraviolet superimposed over continuous spectra.

Tungsten halogen lamps have substantial output in the near infrared. For sources with better coverage of IR-A and IR-B, different sources are more commonly used. Typical outputs from several infrared laboratory sources are shown in Fig. 15. The sources are basically electrical resistance heaters, ceramic and other substances that carry electrical current, which become hot due to ohmic heating, and which emit broadband infrared radiation with moderately high radiance.

In addition to these relatively broadband sources, there are numerous others that emit over more restricted spectral ranges. The light-emitting diode (LED) is one example. It is made of a semiconductor diode with a P-N junction designed so that electrical current through the junction in the forward bias direction produces the emission of optical radiation. The spectral range of emission is limited, but not so much to be considered truly monochromatic. A sample

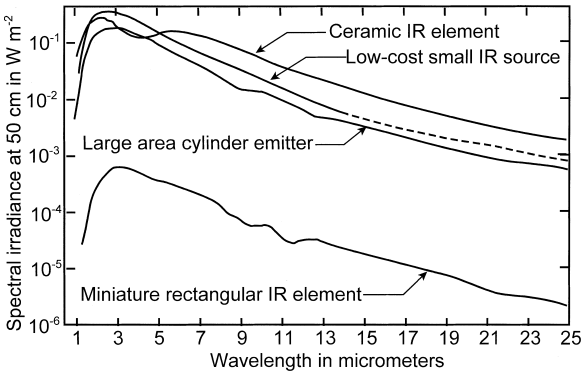


FIGURE 15 Spectral irradiance distributions from four sources of infrared radiation.

LED spectrum is shown in Fig. 16. LEDs are efficient converters of electrical energy into radiant flux.

Lasers deserve special mention. An important characteristic of lasers is their extremely narrow spectral output distribution, effectively monochromatic. A consequence of this is high optical *coherence*, whereby the phases of the oscillations in electric and magnetic field strength vectors are preserved to some degree over time and space. Another characteristic is the high spectral irradiance they can produce. For more information the reader is referred to modern textbooks on lasers and optics. Most gas discharge lasers exhibit a high degree of collimation, an attribute with many useful optical applications.

A problem with highly coherent sources in radiometry and photometry is that not all of the relationships developed so far in this article governing flux levels are strictly correct. The reason is the possibility for constructive and destructive *interference* when two coherent beams of the same wavelength overlap. The superposition of two or more coherent monochromatic beams will produce a combined irradiance at a point that is not always a simple sum of the irradiances of the two beams at the point of superposition. A combined irradiance level can be more and can be less than the sum of the individual beam irradiances, since it depends strongly on the phase difference between the two beams at the point of interest. The predictions of radiometry and photometry can be preserved whenever they are averaged over many variations in the phase difference between the two overlapping beams.

D. Solar Radiation and Daylight

Following its passage through the atmosphere, direct beam solar radiation exhibits the spectral distribution shown in Fig. 17. The fluctuations at wavelengths over 700 nm are the result of absorption by various gaseous constituents of the atmosphere, the most noticeable of which are water vapor and CO_2. The V-lambda curve is also shown in Fig. 17 for comparison.

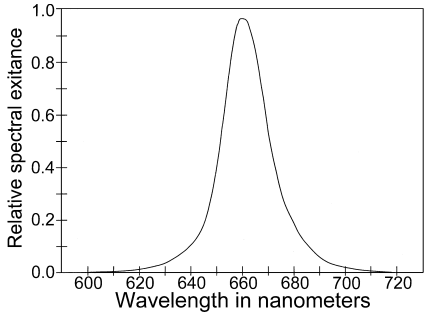


FIGURE 16 Relative spectral exitance of a red light-emitting diode.

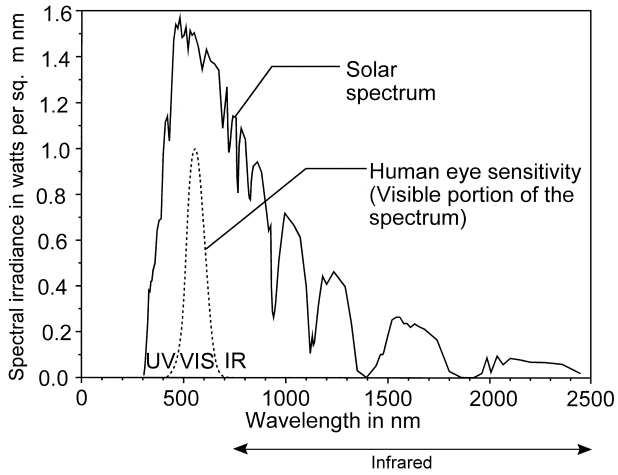


FIGURE 17 Spectral irradiance of terrestrial clear sky direct beam solar radiation.

The spectral distribution of blue sky radiation is similar to that shown in Fig. 17, but the shape is skewed by what is called Rayleigh scattering, the scattering of radiation by molecular-sized particles in the atmosphere. Rayleigh scattering is proportional to the inverse fourth power of the wavelength. Thus, blue light is scattered more prominently than red light. This is responsible for the blue appearance of sky light. (The accompanying removal of light at short wavelengths shifts the apparent color of beam sunlight toward the red end of the spectrum, responsible for the orange-red appearance of the sun at sunrise and sunset.) The spectral distribution of daylight is important in the field of colorimetery and for many other applications, including the daylight illumination of building interiors.

VII. OPTICAL PROPERTIES OF MATERIALS

A. Introduction

Central to radiometry and photometry is the interaction of radiation with matter. This section provides a discussion of the properties of real materials and their abilities to emit, reflect, refract, absorb, transmit, and scatter radiation. Only the rudiments can be addressed here, dealing mostly with terminology and basic concepts. For more information on the optical properties of matter, the reader is directed to available texts on optics and optical engineering, as well as other literature on material properties.

B. Terminology

The improved uniformity in symbols, units, and nomenclature in radiometry and photometry has been extended to the optical properties of materials. Proper terminology can now be identified for the processes of reflection, transmission, and emission of radiant flux by or through material media. Although symbols have been standardized for most of these properties, there are a few exceptions.

To begin, the CIE definitions for *reflectance*, *transmittance*, and *absorptance* are provided:

1. *Reflectance* (for incident radiation of a given spectral composition, polarization and geometrical distribution) (ρ): Ratio of the reflected radiant or luminous flux to the incident flux in the given conditions (unit: 1)
2. *Transmittance* (for incident radiation of given spectral composition, polarization and geometrical distribution) (τ): Ratio of the transmitted radiant or luminous flux to the incident flux in the given conditions (unit: 1)
3. *Absorptance*: Ratio of the absorbed radiant or luminous flux to the incident flux under specified conditions (unit: 1)

These definitions make explicit the point that radiation incident upon a surface can have nonconstant distributions over the directions of incidence, over polarization state, and over wavelength (or frequency). Thus, when one wishes to measure these optical properties, it must be specified how the incident radiation is distributed in wavelength and direction and how the emergent detected radiation is so distributed if the measurement is to have meaning. Polarization effects are not dealt with here. The wavelength dependence of radiometric properties of materials is indicated with a functional lambda λ thus: $\tau(\lambda)$, $\rho(\lambda)$, and $\alpha(\lambda)$.

The directional dependencies are indicated by specifying the spherical angular coordinates (θ, ϕ) of the incident and emergent beams.

In other fields it is common to assign the ending *-ivity* to *intensive*, inherent, or bulk properties of materials. The ending *-ance* is reserved for the *extensive* properties of a fixed quantity of substance, for example a portion of the substance having a certain length or thickness. (Sometimes the term *intrinsic* is used instead of intensive and *extrinsic* is used instead of extensive.) Figure 18 illustrates the difference between intrinsic and extrinsic reflection properties and introduces the concept of interface reflectivity. An example from electronics is the 30 ohm electrical resistance of a 3 cm length of a conductor having a *resistivity* of 10 ohms/cm.

According to this usage in radiometry, *reflectance* is reserved for the fraction of incident flux reflected (under defined conditions of irradiation and reception) from a finite and specified portion of material, such as a 1-cm-thick plate of fused silica glass having parallel, roughened

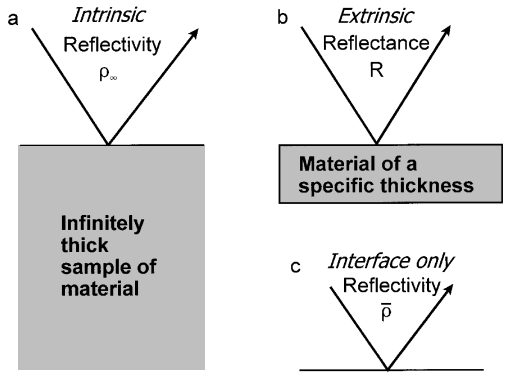


FIGURE 18 (a) Intrinsic versus (b) extrinsic reflection properties of a material. (c) Interface reflectivity.

surfaces in air. The *reflectivity* of a material, such as BK7 glass, would refer to the ratio of reflected to incident flux for the perfectly smooth (polished) interface between an optically infinite thickness of the material and some other material, such as air or vacuum. The infinite thickness is specified to ensure that reflected flux from no other interface can contribute to that reflected by the interface of interest, and to ensure the collection of subsurface flux scattered by molecules of the material.

CIE definitions for the intrinsic optical properties of matter read as follows:

1. *Reflectivity* (of a material) (ρ_∞): Reflectance of a layer of the material of such a thickness that there is no change of reflectance with increase in thickness (unit: 1)
2. *Spectral transmissivity* (of an absorbing material) ($\tau_{i,o}(\lambda)$): Spectral internal transmittance of a layer of the material such that the path of the radiation is of unit length, and under conditions in which the boundary of the material has no influence (unit: 1)
3. *Spectral absorptivity* (of an absorbing material) ($\alpha_{i,o}(\lambda)$): Spectral internal absorptance of a layer of the material such that the path of the radiation is of unit length, and under conditions in which the boundary of the material has no influence (unit: 1)

One can further split the reflectivity ρ_∞ into interface-only and bulk property components. We use the symbol $\bar{\rho}$, *rho* with a bar over it, to indicate the interface contribution to the reflectivity. The interface transmissivity $\bar{\tau}$ is included in this notational custom. Since there is presumed to be no absorption when radiation passes through or reflects from an interface, $\bar{\tau} + \bar{\rho} = 1.0$.

To denote the optical properties of whole objects, such as parallel sided plates of a material of specific thickness, we use upper case Roman font characters, as with the symbol R for reflectance, and the "-ance" suffix.

This terminology is summarized as follows:

$\bar{\rho}$ Reflectivity of an interface
ρ Reflectivity of a pure substance, including both bulk and interface processes
R Reflectance of an object
$\bar{\tau}$ Transmissivity of an interface
τ (Internal) linear transmissivity of (a unit length of) a transparent or partially transparent substance, away from interfaces; unit: m^{-1}
T Transmittance of an object
α (Internal) linear absorptivity of (a unit length of) a transparent or partially transparent substance, away from interfaces; unit: m^{-1}
A Absorptance of an object

C. Surface and Interface Optical Properties

1. Conductor Optical Properties

A perfect conductor, characterized by infinitely great conductivity, has an infinite refractive index and penetration of electromagnetic radiation to any depth is prohibited. This produces perfect reflectivity. Real conductors such as aluminum and silver do not have perfect conductivities nor do they have perfect reflectivities. Their reflectivities are quite high, however, over broad spectral ranges. They are therefore useful in radiometric and photometric applications. Unprotected mirrors made of these materials, unfortunately, tend to degrade with exposure to air over time and they are seldom used without protective overcoatings. The normal incidence spectral reflectances of optical quality glass mirrors coated with aluminum, with aluminum having a magnesium fluoride protective overcoat, with aluminum having a silicon monoxide overcoat, with silver having a protective dielectric coating, and with gold are shown in Fig. 19. The reflectance of these surfaces, already quite high at visible and infrared wavelengths, increases with incidence angle, approaching unity at 90°.

2. Nonconductor Optical Properties

Consider the extremely thin surface region of a perfectly smooth homogeneous and isotropic dielectric material, its interface with another medium such as air, water, or a vacuum, an interface normally too thin to absorb significant quantities of the radiation incident on it. Absorption is not considered in this discussion since it is considered to be a bulk or volume characteristic of the material. Radiation incident upon an interface between two different materials is split into two parts. Some is reflected, and the rest is transmitted. The fraction of incident flux that is reflected is called the interface reflectivity $\bar{\rho}$ and the fraction

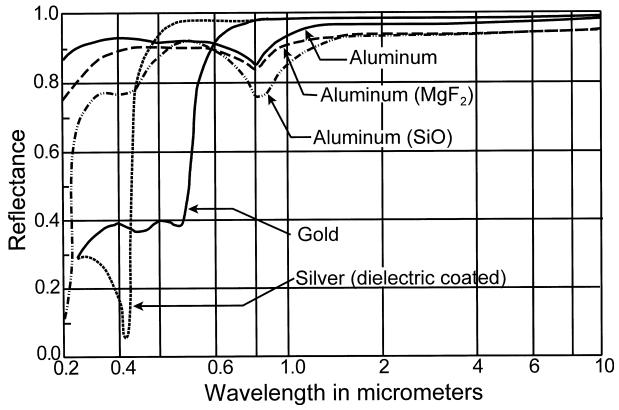


FIGURE 19 Spectral reflectances of commercially available metallic mirror materials.

transmitted is the interface transmissivity $\bar{\tau}$. The variations of $\bar{\rho}$ and $\bar{\tau}$ with angle of incidence are given by Fresnel's formulas, which can be found in most optical textbooks.

When the bulk medium optical properties are considered, the situation is more complicated, since the transmitted flux can be absorbed and "re-reflected" and/or scattered by the medium below the interface, by direction- and wavelength-dependent processes.

When both interface and interior optical processes are considered together, the spectral and directional variations in transmissivity and reflectivity become still more important, and the absorptivity of the medium also comes into play. The wavelength dependence of the optical properties of materials is indicated with a functional λ notation thus: $\tau(\lambda)$, $\alpha(\lambda)$, and $\rho(\lambda)$. The direction of an element of solid angle $d\omega$ is indicated using the spherical angular coordinates (θ, ϕ), illustrated in Fig. 6. Using these coordinates, the directional dependence of optical properties is indicated with the functional notation: $\tau(\theta, \phi)$ and $\rho(\theta, \phi)$, and the combined spectral and directional properties thus: $\tau(\lambda, \theta, \phi)$ and $\rho(\lambda, \theta, \phi)$.

3. Surface Emission Properties

The emissive properties of greybody and nonblackbody surfaces are characterized by their *emissivity* ϵ. Emissivity is the ratio of the actual emission of thermal radiant flux from a surface to the flux that would be emitted by a perfect blackbody emitter at the same temperature. According to the terminology guidelines given earlier, the term *emissivity* should be reserved for the surface of an infinitely thick slab of pure material with a polished surface, while emittance would apply to a finite thickness of an actual object. For substances opaque at the wavelengths of emission, however, the intrinsic and extrinsic versions of ϵ are the same, leading to two acceptable names for the same quantity.

As was the case for reflectance and transmittance, *emittance* is in general a directional quantity and can be specified as $\epsilon(\theta, \phi)$. The directional emittance at normal incidence ($\theta = 0$) is called the *normal emittance*. The average of the directional emittance over the whole hemispherical solid angle is called the *hemispherical emittance*. The emittances shown in Table VII are for hemispherical emittance into a vacuum.

The spectral exitance $M_\lambda(\lambda)$ of a nonblackbody can be specified using the spectral emittance $\epsilon(\lambda)$:

$$M_\lambda(\lambda) = \epsilon(\lambda) M_{\mathrm{bb}\lambda}(\lambda) \tag{41}$$

with $M_{\mathrm{bb}\lambda}$ being given by (40).

4. Directional Optical Properties

Radiation incident at a point in a surface can come to that point from many directions. The concept of a pencil of rays, rays filling a right circular conical solid angle, like the shape of the tip of a well-sharpened wooden pencil, is useful in describing the directional dependences of transmittance and reflectance, for both theoretical treatments and in practical measurements.

In making transmittance or reflectance measurements, a sample to be tested is illuminated with radiation filling some solid angular range of directions. The reflected or transmitted flux is then collected over another range of directions within some second solid angle.

In order for the transmittance or reflectance value to have meaning, either theoretically or experimentally, it is essential that the directions and solid angles of incidence and emergence be specified. These tell the ranges of angles involved in the measurement.

In discussing reflectance and transmittance, there are three categories of solid angles of interest, and several different definitions of reflectance and transmittance using combinations of these. The three solid angle categories are *directional*, *conical*, and *hemispherical.* They are illustrated in Fig. 20. There are nine possible combinations

TABLE VII Hemispherical Emittance Values for Typical Materials

Material	Emittance from 4 to 16 μm
White paint	0.90
Black asphalt and roofing tar	0.93
Light concrete	0.88
Pine wood	0.60
Stainless steel	0.18 to 0.28
Galvanized sheet metal	0.13 to 0.28
Aluminum sheet metal	0.09
Polished aluminum	0.05 to 0.08

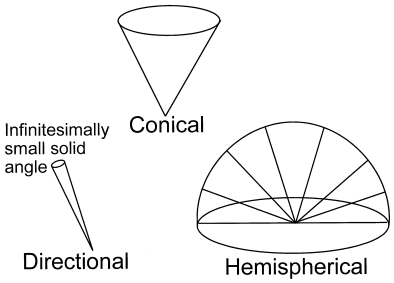


FIGURE 20 Geometry for directional, conical, and hemispherical solid angles.

of these three kinds of solid angle, resulting in the nine names for them given below. The most commonly used ones are indicated in bold face type.

- **Bidirectional**
- Directional–conical
- Directional–hemispherical
- Conical–directional
- Hemispherical–directional
- **Biconical**
- **Conical–hemispherical**
- Hemispherical–conical
- **Bihemispherical**

The first five of these are mainly found in theoretical discussions. The last four are used in reflectance and transmittance measurements. Solar optical property standards published by various organizations refer to conical–hemispherical measurements. The reason is that for most practical problems, it is only the total transmitted or reflected irradiance due to the directly incident beam alone that is of interest. For other applications and more general or more complex situations, the biconical definition is the most important (see Fig. 21). Theoretical treatments

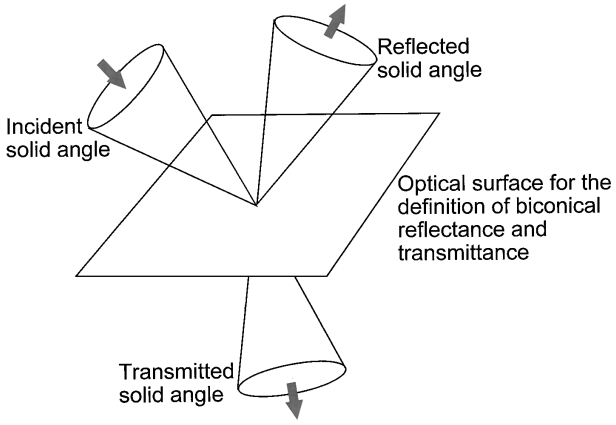


FIGURE 21 Geometry for the definition of biconical transmittance and reflectance.

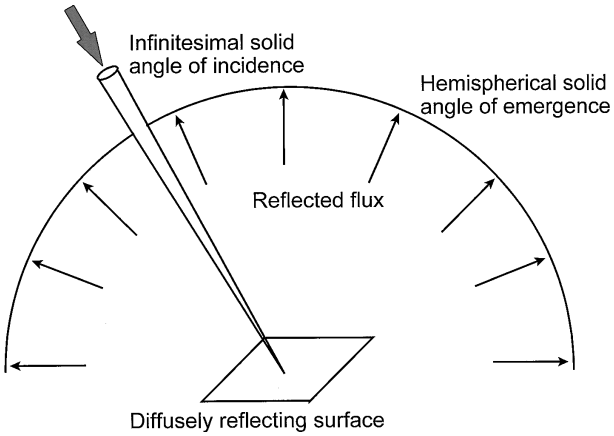


FIGURE 22 Geometry for the definition of directional–hemispherical reflectance.

of radiative transfer deal almost exclusively with the "directional" versions of the definitions. Sometimes the terminology "bidirectional" is used to refer to biconical measurements. This is appropriate when the solid angles involved are small. Example geometries are shown in Figs. 22 and 23.

VIII. THE DETECTION OF RADIATION

There is considerable variety in the kinds of devices (called *detectors* or *sensors*) available for the detection and measurement of optical radiation. Some respond to the heat produced when radiant energy is absorbed by a surface. Some convert this heat into mechanical movement, and some convert it into electricity. Photographic emulsions convert incident radiation into chemical changes

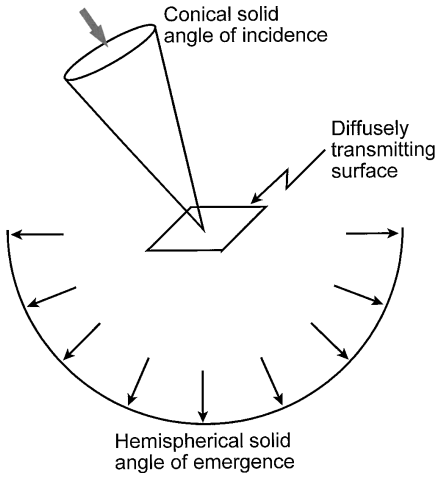


FIGURE 23 Geometry for the definition of conical–hemispherical transmittance.

made visible by the development process. Other detectors convert electromagnetic radiation directly into electrical energy.

Many electrical effects have been devised to amplify the typically small electrical signals produced by detectors to levels easier to measure. There are unavoidable small fluctuations found in the output signals of all detectors which mask or obscure the signal resulting from incident radiation. This is called *noise*, and various means have been devised to reduce its effect on measurement results.

Most detectors with high sensitivity (strong response to weak flux levels) have nonuniform spectral responses. Often the inherent spectral response of the detector is not the one desired for the application. In most such cases it is possible to add a spectrally selective filter, producing a combined filter/detector response closer to what is desired. Matching filters with detectors for this purpose can be difficult, but it is one of the most important problems in radiometry and photometry.

The output voltage or current of most detectors depends on more than just the strength of the incident flux. Temperature T can have an effect, as can the direction θ of incident radiation. If we combine all these dependencies into one single spectral response function, $R(\lambda, \Phi_\lambda, T, \theta, x, y, z, \ldots)$, we can write an equation for the output signal $S(\lambda)$ at wavelength λ as a function of the incident spectral radiant flux Φ_λ.

$$S(\lambda) = R(\lambda, \Phi_\lambda, T, \theta, x, y, z, \ldots)\,\Phi_\lambda + S_o, \qquad (42)$$

where x, y, and z are other physical parameters on which the detector's output might depend and S_o is the "dark signal," the signal output of the detector (be it current or voltage) when the flux on it is zero.

Considering only the spectral dependency in the above equation, the total output signal S, in terms of the incident spectral irradiance $E_\lambda(\lambda)$, a spectral altering filter transmittance $T(\lambda)$, the detector responsivity $R(\lambda)$, and the detector area A will be given by

$$S = A\int_0^\infty E_\lambda(\lambda)\,T(\lambda)\,R(\lambda)\,d\lambda + S_o. \qquad (43)$$

If the detector spectral response $R(\lambda)$ is constant, at the value R_o, over some wavelength range of interest, a spectrum altering filter will not be needed and the only integral remaining in (43) is over the spectral irradiance. The result is the simpler equation

$$S = AE_eR_o + S_o, \qquad (44)$$

which may be solved for the incident irradiance.

$$E_e = k(S - S_o), \qquad (45)$$

where $k = 1/AR_o$ is the calibration constant for a detector used as a normal incidence irradiance meter and S_o is the *dark signal*.

Values are published by manufacturers for the responsivity and other characteristics of their detectors. These performance figures are approximate and are used mainly for the selection of a detector with the proper characteristics for the given application—not for calibration purposes, with a notable exception, described in Section X.C.

It is important to note that the smaller the detector, the lower the noise level produced. There is therefore usually a noise penalty for using a detector having a sensitive surface significantly larger than the incident beam. The unused area contributes to both the dark current and to the noise but not to the signal. The signal-to-noise ratio (SNR) of a detector can therefore be improved by using a detector only as large as needed to match the beam of flux placed on the detector by the conditioning optics.

Often the flux incident on a detector is "chopped," is made to switch on and off at some frequency f. Much of the noise in such detectors can be suppressed from the output signal if the alternating output signal from the detector is amplified only at the chopping frequency f. The larger the frequency bandwidth Δf of this amplification circuit, the greater the noise in the amplified signal. This leads to the concept of *noise equivalent power*, or NEP, of the detector. This is the flux incident on the detector, in units of watts, which produces an amplified signal just equal to the *root mean square* (rms) of the noise. It is generally desirable to have a low value of the NEP, which is quoted in units of $W \cdot Hz^{-1/2}$. The lower the value of the NEP the lower the flux the detector can measure with a good SNR. *Detectivity D* is the reciprocal of NEP. *Normalized detectivity, D**, is the detectivity normalized for detector area and frequency bandwidth. It has units of $Hz^{\frac{1}{2}} \cdot (cm^2)^{1/2} \cdot W^{-1}$.

The spectral detectivities of a variety of detectors are shown in Fig. 24. One thing is clear from those plots. Broad spectral coverage generally comes at the expense of detectivity.

Once an appropriate detector has been selected, it will generally be placed in an optical–mechanical system having the effect of conditioning the flux prior to its receipt by the detector. This conditioning can consist of chopping, focusing incident flux into a narrow conical solid angular range of angles, and/or spectral filtering.

IX. RADIOMETERS AND PHOTOMETERS, SPECTRORADIOMETERS, AND SPECTROPHOTOMETERS

A. Introduction

Radiometer is the term given to an instrument designed to measure radiant flux. Some radiometers measure the radiant flux contained in a beam having a known solid

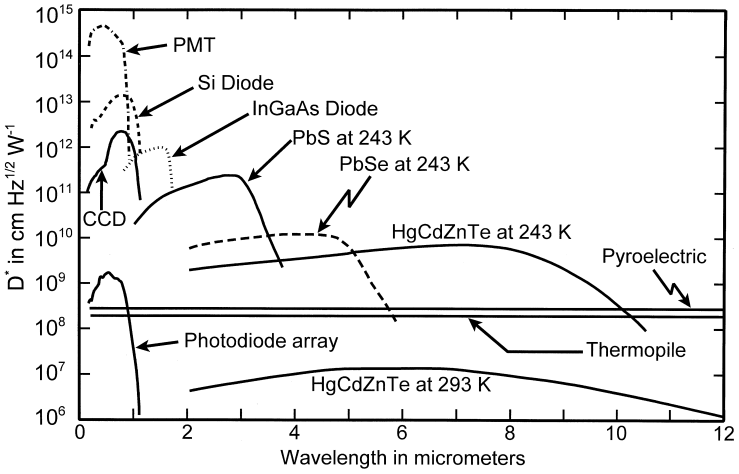


FIGURE 24 Representative spectral normalized detectivities of a variety of detectors.

angle and cross-sectional area. Others measure the flux received from a large range of solid angles. Some radiometers measure over a large wavelength range. These are termed *broadband*. Others perform measurements only over a narrow spectral interval. When the shape of the spectral response of a broadband radiometer is made to match the human spectral photopic efficiency function, the V-lambda curve, it is called a *photometer.*

Some narrow spectral interval radiometers are made which scan the position of their narrow spectral interval across the spectrum. These are called *spectroradiometers.* They are used to measure the spectral flux, irradiance, or radiance received by them.

Spectrophotometers are misnamed. This term is generally applied to neither radiometers nor photometers but to transmissometers or to reflectometers—instruments measuring an optical property—which scan over a range of monochromatic wavelengths. In spite of the inclusion of "photo" in the name, the human photopic spectral response function (the V-lambda curve) is generally not employed in the use of spectrophotometers. They might therefore more properly be called *spectral transmissometers* (or *reflectometers*).

Radiometers are divided into radiance and irradiance subclasses. Instruments with intentionally broad spectral coverage are called *broadband radiometers.* Photometers are similarly divided into luminance and illuminance versions.

B. Spectral Response Considerations

In the practical use of radiance (and irradiance) meters, it is especially important to be cognizant of the spectral limitations of the meter and to include these limits when reporting measurement results. Flux entering the meter having wavelengths outside its range of sensitivity will not be measured. In such cases, the measurements will only sample a portion of the incident flux and should be so reported. Well-built photometers do not share this characteristic. If the spectral response of a photometer strictly matches the shape of the V-lambda curve, then flux outside the visible wavelength range *should* not be measured, will not be measured, will not be recorded, and cannot be reported. Furthermore, in this case of perfect spectral correction, *any* spectral distribution of radiation incident on the photometer will be measured correctly without spectral response errors. On the other hand, if a photometer's spectral response does *not* quite match the shape of the V-lambda curve, the resulting errors can be small or large depending upon the spectral distribution of flux from the source over the spectral region of the departure from $V(\lambda)$ response. Consider, for example, the case of a measurement of the illuminance from a Helium Neon laser beam at wavelength 632.8 nm. This wavelength is at the red edge of the visible spectrum and a relatively small error in the V-lambda correction of a photometer at this wavelength can yield a large error in the measurement of illuminance from this source.

C. Cosine Correction

A consequence of the cosine law is that the output of a perfect irradiance meter illuminated uniformly with collimated radiation fully filling its sensing area will decrease with the cosine of the angle of incidence as that angle increases from zero to 90°. Such behavior is called "good cosine correction." Most detectors do not have this desirable characteristic by themselves. To restore good cosine response, some correction method is needed if a detector is to work properly as an irradiance or illuminance meter. Furthermore, the housings of many detectors shade their sensitive surfaces at some angles of incidence, again calling for some means of angular response correction.

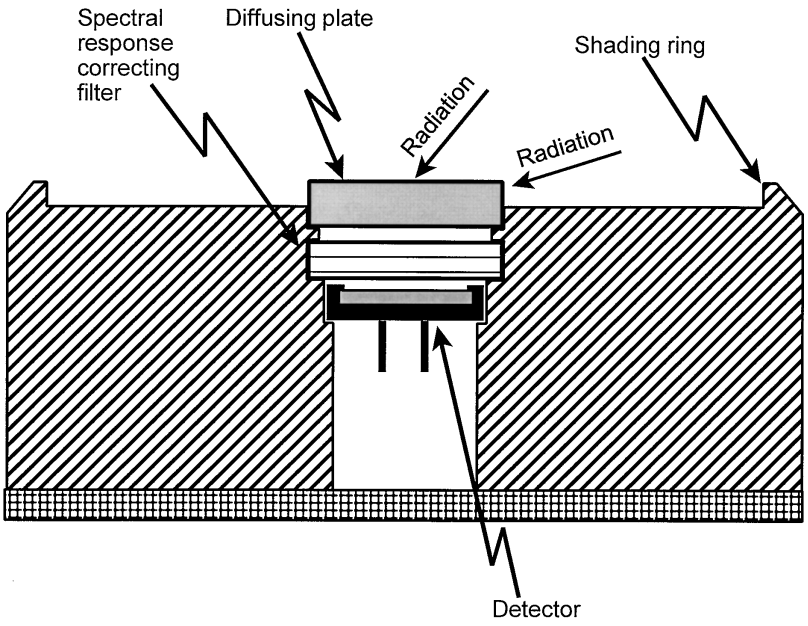


FIGURE 25 Schematic illustration of the features of an irradiance/illuminance meter.

A common method, shown in Fig. 25, for providing the needed correction is to cover the detector with a sheet of milk-white, highly diffusing, semitransparent material having good (and ideally constant) hemispherical–conical spectral transmittance over the spectral range of good detector sensitivity. The idea is that no matter how the incident radiation falls on this material, a fixed and constant fraction of it will be delivered to the detector over a range of angles. In practice, no diffusing sheet has been found that satisfies this ideal perfectly.

What is done is to experiment with a variety of diffusing materials, surface roughnesses, and geometrical configurations until a combination is found that provides reasonably good cosine correction.

A solution to this problem is to limit the size of the diffusing sheet and allow it to extend above the detector housing so that at large angles of incidence some of the incident flux will be received by the *edge* of the sheet, this edge being perpendicular to the front surface. Thus, as more and more flux is reflected from the front of the sheet, more and more will be transmitted through its edge, since in this case the incidence angle is decreasing and the exposed area incrases. At an 80° angle of incidence, for example, little flux will enter through the front face of the diffuser, but much more will enter through the edge.

A problem remains, however. True cosine response drops to zero at 90°, whereas the exposed edge of the diffuser receives considerable quantities of flux at this angle. The cosine corrector must be designed so that no flux can reach the detector for angles of incidence at and greater than 90°. The usual solution to this requirement is illustrated generically in Fig. 25.

As the angle of incidence increases, more and more flux will reach the edge of the detector until the angle of incidence approaches 90°, at which point the shading ring begins to shade the edge. Finally, at 90°, the diffuser is shaded completely and no flux can reach it. The design of the specific dimensions of this cosine correction scheme depends strongly on the biconical optical properties of the diffusing sheet, the geometrical placement of the detector below it, and the angular response characteristics of the detector itself. Finding the right geometry is often a hit-or-miss proposition. Even if a good design is found, the quality of the corrected cosine response can suffer if the properties of the diffusing material change in time, from batch to batch of manufacture, or with the wavelength of incident radiation. Making a good cosine corrector is one of the most difficult problems in the manufacture of good quality, accurate irradiance and illuminance meters.

A way of providing better cosine correction than the one diagramed in Fig. 25 is through the use of an integrating sphere. An arrangement for utilizing the desirable properties of the integrating sphere is illustrated in Fig. 26. The approach is based on the following idealized principle. Flux entering a small port in a hollow sphere whose interior surface is coated with a material of extremely high diffuse reflectance will be multiply reflected (scattered) a large number of times, in all directions, with little loss on each reflection. If a small hole is placed in the side of the sphere and shielded from flux that has not been reflected at least once by the sphere, the flux emerging from this hole

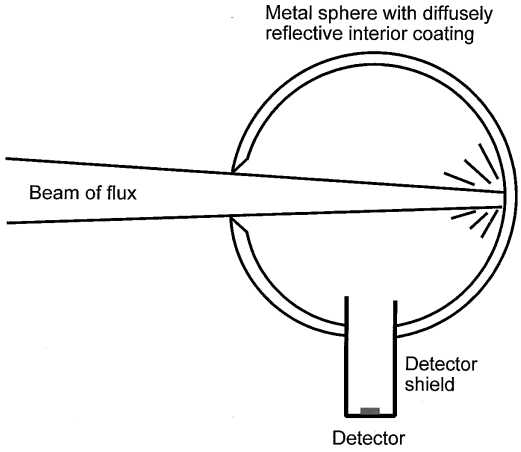


FIGURE 26 Integrating sphere cosine correction.

will be a fixed and constant fraction of the flux entering the other hole, regardless how the flux entering the input port is distributed in angle.

Real integrating spheres, with reflectances less than 1.0 and with entrance and exit ports of finite areas cannot achieve this idealized performance. They can be made to approach it closely, but they are not efficient at delivering flux to the detector for measurement, so irradiance and illuminance meters employing integrating spheres generally suffer lower sensitivities.

X. CALIBRATION OF RADIOMETERS AND PHOTOMETERS

A. Introduction

Radiometers and photometers involve a number of components, all contributing to the overall sensitivity of the instrument to incident radiation. Although one could in principle determine the contribution of each individual component to the overall calibration of the instrument, in practice this procedure is seldom used. Instead, the complete instrument is calibrated all at once.

Calibration is usually a two-step process. First one determines the mathematical transformation needed to convert an output electrical signal into an estimate of the input flux in the units desired for the quantity being measured. Second, one ensures the accuracy of this transformation over time as the characteristics of the components making up the radiometer or photometer change or drift.

There are two approaches to calibrating or recalibrating a radiometer/photometer. In the first case, one uses the radiometer/photometer to measure flux from a standard source whose emitted flux is known accurately in the desired units and then applies a suitable transformation to convert the output signal to the proper magnitude and units of the standard input. For this to work, it is critical that the overall response of the radiometer/photometer be constant over the period of time between calibrations. The output conversion transformation can be either in hardware, where the sensitivity of the radiometer is adjusted so that it reads correctly, or in "software," where a calibration constant is multiplied by the output signal to convert it to the proper value and units every time a measurement is made.

In the second approach to calibration, one measures the flux from an uncalibrated source, first with the device to be calibrated, and then with an already-calibrated standard radiometer/photometer having identical field of view and spectral response. The output of the device is then calibrated to be identical to the measured result and units obtained with the standard radiometer/photometer. Once the calibration is performed, or the calibration transformation is known, it can be applied to subsequent measurements and the device is thereby said to be calibrated.

B. Standard Sources

For radiometers and photometers whose calibration drifts slowly over time, one can calibrate the device when it is first fabricated and then recalibrate it periodically over some acceptable time period. For most accurate results, it is advisable to recalibrate frequently at first, and to then increase the time interval between recalibrations only after a history of drift has been established. For precise radiometry and photometry, a *working standard* or *transfer standard* is used to make frequent calibration checks between (or even during) measurements to account for the effects of small residual drifts in the calibration of a radiometer or photometer.

Historically, the focus of calibration was on the preparation of standard sources, most notably standard lamps, which produce a known and constant quantity of flux giving a known irradiance at a fixed distance from the emitting element. A typical measurement configuration is illustrated schematically in Fig. 27(a). There has been a shift to the use of calibrated detection standards; that is, detectors whose responsivity is sufficiently constant and reproducible over time to make it possible to calibrate other detectors or radiometers based on these standard detectors. Standard lamps are still available as calibrated sources, however. These generally produce a fixed output flux distribution with wavelength. Because of the possibility of nonlinear response effects, radiometers and photometers should be calibrated only over their ranges of linearity, within which a standard lamp can be found.

Many nations maintain primary standards for radiometry (and photometry) in national laboratories dedicated to this purpose. In the United States, such standards are maintained by the National Institute of Standards and

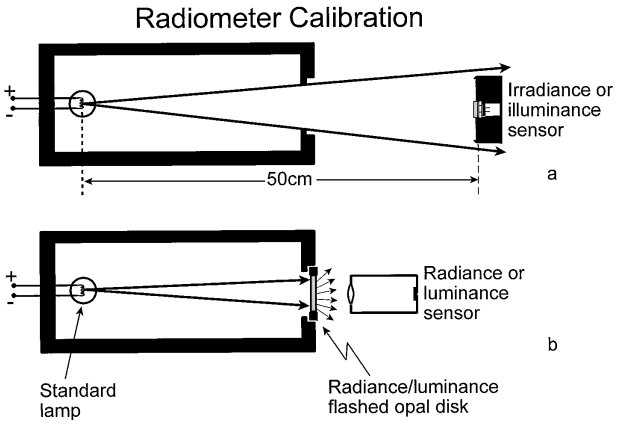


FIGURE 27 (a) Calibration arrangement for irradiance/illuminance. (b) Calibration arrangement for radiance/luminance.

Technology in Gaithersburg, MD. From these are derived secondary standards (also called transfer standards) that can be maintained at private laboratories or by other organizations for the purpose of calibrating and recalibrating commercially and custom produced radiometers and photometers. Working standards are standards derived from secondary standards but which are designed and intended for easy and repeated use to check the calibration of a radiometric or photometric system periodically during or between measurements.

1. Calibration of Radiance and Irradiance Meters

Calibrations using a standard lamp frequently utilize specially designed tungsten filament lamps whose emitting characteristics are known to be quite constant over a period of time if the lamp is not frequently used. Such lamps must be operated with precisely the same electrical current through the filament as when their calibrations were initially set. Specially designed power supplies are made for use with such lamps. These power supplies ensure the constancy of this filament current and also keep track of how many hours the filament has been operated since initial calibration.

One can obtain irradiance standard lamps commercially and use them for the calibration of broadband irradiance sensors. They must be operated according to manufacturer specifications and care must be taken to avoid stray light from the source reflecting from adjacent objects and into the radiometer being calibrated.

Over the years, researchers at the National Institute of Standards and Technology have worked to develop improved standards of spectral radiance and irradiance for the ultraviolet, visible, and near infrared portions of the spectrum. The publications of that U.S. government agency should be consulted for the details.

2. Calibration of Luminance and Illuminance Meters

Standard sources of radiance and irradiance that emit usable quantities of radiation over the visible portion of the spectrum can be used as standards for the calibration of photometers if the photometric outputs of these sources is known. Commercial radiometric and photometric standards laboratories generally can supply photometric calibrations for their radiometric sources for modest additional cost. The most common source is the incandescent filament lamp, with its characteristic spectral output distribution. If the primary use of the photometer being calibrated is to measure light levels derived from sources with similar spectral distributions, and if the V-lambda correction of the photometer is good, then use of tungsten filament standard lamps is an acceptable means of calibration.

If the photometer is intended for measurement of radiation with substantially different spectral distribution and the V-lambda correction is not good, then significant measurement errors can result from calibration using tungsten sources. Fortunately, other standard spectral distributions have been defined. They are based on phases of daylight (primarily for colorimetric applications). Sources exhibiting approximations of these distributions have been developed. For cases of imperfect V-lambda correction, it is recommended that calibration sources be used that more closely match the distributions to be measured with the photometer.

C. Calibrated Detectors

Calibrated silicon photodetectors are now available as transfer or working standards based on the NIST *absolute spectral responsivity scale.* Current information about NIST calibration services can be found at web site http://www.physics.nist.gov.

D. National Standards Laboratories

Anyone concerned with calibration of radiometers and photometers can benefit greatly from the work of the National Institute of Standards and Technology and its counterparts in other countries.

NSSN offers a web-based comprehensive data network on national, foreign, regional, and international standards and regulatory documents. A cooperative partnership between the American National Standards Institute (ANSI), U.S. private-sector standards organizations, government agencies, and international standards organizations, NSSN can help in the identification and location of national standards laboratories offering services in radiometry and photometry outside the United States.

World Wide Web address: http://www.nssn.org. The web address for the International Standards Organization (ISO) is http://www.iso.ch. A list of national metrology laboratories can be found at http://www.vnist.gov/oiaa/national.htm.

ACKNOWLEDGMENT

Portions reprinted with permission from McCluney, R. (1994). "Introduction to Radiometry and Photometry," Artech House, Inc., Norwood, MA. www.artechhouse.com.

SEE ALSO THE FOLLOWING ARTICLES

COLOR SCIENCE • INFRARED SPECTROSCOPY • LIGHT SOURCES • OPTICAL DETECTORS • POLARIZATION AND POLARIMETRY • PHOTONIC BANDGAP MATERIALS • RADIATION, ATMOSPHERIC • RADIATION EFFECTS IN ELECTRONIC MATERIALS AND DEVICES • RADIATION SOURCES • RADIO ASTRONOMY, PLANETARY • REMOTE SENSING FROM SATELLITES

BIBLIOGRAPHY

Biberman, L. M. (1967). "Apples Oranges and UnLumens," *Appl. 0ptics* **6,** 1127.

Boyd, R. W. (1983). "Radiometry and the Detection of Optical Radiation," Wiley, New York.

Budde, W. (1983). "Optical Radiation Measurements," Wiley, New York.

Chandrasekhar, S. (1960). "Radiative Transfer," Dover Publications, New York.

CIE (1990). "CIE 1988 2º Spectral Luminous Efficiency Function for Photopic Vision," Tech. Rept. CIE 86. CIE, Vienna, Austria.

CIE, (1987). "International Lighting Vocabulary," 4th ed., Publ. No. 17.4. Commission International de l'Eclairage (CIE), Vienna, and International Electrotechnical Commission (IEC). [Available in the U.S. from TLA-Lighting Consultants, 7 Pond St., Salem, MA 01970.

Dereniak, E. L., and Crowe, D. G. (1984). "Optical Radiation Detectors," Wiley, New York.

Goebel, D. G. (1967). Generalized integrating sphere theory. *Appl. Optics* **6,** 125–128.

Grum F., and Becherer, R. J. (1979). "Optical Radiation Measurements. Volume 1 Radiometry," Academic Press, New York.

IES (2000). "The IESNA Lighting Handbook: Reference and Applicaiton," 9th ed., Illuminating Engineering Society of North America, New York.

McCluney, R. (1994). "Introduction to Radiometry and Photometry," Artech House, Norwood, MA.

Meyer-Arendt, J. R. (1968). Radiometry and photometry: units and conversion factors. *Appl. Optics* **7,** 2081–2084.

Nicodemus, F. E. (1963). Radiance. *Am. J. Phys.* **31,** 368–377.

Nicodemus, F. E. (1976). "Self-Study Manual on Optical Radiation Measurements," NBS Technical Note 910, U.S. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.

Siegel, R., and Howell, J. R. (1992). "Thermal Radiation Heat Transfer," 3rd ed., Hemispherical Publishing/McGraw-Hill, New York.

Spiro, I. J., and Schlessinger, M. (1989). "Infrared Technology Fundamentals," Marcel Dekker, New York.

Taylor, B. N. (1995). "Guide for the Use of the International System of Units (SI)," NIST Special Publication 811, National Institute of Standards and Technology, Gaithersburg, MD.

Welford, W. T., and Winston, R. (1989). "High Collection Nonimaging Optics," Academic Press, New York.



Superstring Theory

John H. Schwarz
California Institute of Technology

I. Supersymmetry
II. String Theory Basics
III. Superstrings
IV. From Superstrings to M-Theory

GLOSSARY

Compactification The process by which extra spatial dimensions form a very small (compact) manifold and become invisible at low energies. To end up with four large dimensions, this manifold should have six dimensions in the case of superstring theory or seven dimensions in the case of M theory.

D-brane A special type of p-brane that has the property that a fundamental string can terminate on it. Mathematically, this corresponds to Dirichlet boundary conditions, which is the reason for the use of the letter D.

M-theory A conjectured quantum theory in eleven dimensions, which is approximated at low energies by eleven-dimensional supergravity. It arises as the strong coupling limit of the type IIA and E8 $\times$ E8 heterotic string theory. The letter M stands for magic, mystery, or membrane according to taste.

p-brane A dynamical excitation in a string theory that has p spatial dimensions. The fundamental string, for example, is a 1-brane. All of the other p-branes have tensions that diverge at weak coupling, and therefore they are nonperturbative.

S duality An equivalence between two string theories (such as type I and $SO(32)$ heterotic) which relates one at weak coupling to the other at strong coupling and vice versa.

String theory A relativistic quantum theory in which the fundamental objects are one-dimensional loops called strings. Unlike quantum field theories based on point particles, consistent string theories unify gravity with the other forces.

Supergravity A supersymmetric theory of gravity. In addition to a spacetime metric field that describes spin 2 gravitons, the quanta of gravity, these theories contain one or more spin $3/2$ gravitino fields. The gravitino fields are gauge fields for local supersymmetry.

Superstring A supersymmetric string theory. At weak coupling there are five distinct superstring theories, each of which requires ten-dimensional spacetime (nine spatial dimensions and one time dimension). These five theories are related by various dualities, which imply that they are different limits of a single underlying theory.

Supersymmetry A special kind of symmetry that relates bosons (particles with integer intrinsic spin) to fermions (particles with half-integer intrinsic spin). Unlike other symmetries, the associated conserved charges transform as spinors. According to a

fundamental theorem, supersymmetry is the unique possibility for a nontrivial extension of the known symmetries of spacetime (translations, rotations, and Lorentz transformations).

T duality An equivalence between two string theories (such as type IIA and type IIB) which relates one with a small circular spatial dimension to the other with a large circular spatial dimension and vice versa.

MANY of the major developments in fundamental physics of the past century arose from identifying and overcoming contradictions between existing ideas. For example, the incompatibility of Maxwell's equations and Galilean invariance led Einstein to propose the special theory of relativity. Similarly, the inconsistency of special relativity with Newtonian gravity led him to develop a new theory of gravity, which he called the general theory of relativity. More recently, the reconciliation of special relativity with quantum mechanics led to the development of quantum field theory. We are now facing another crisis of the same character. Namely, general relativity appears to be incompatible with quantum field theory. Any straightforward attempt to "quantize" general relativity leads to a nonrenormalizable theory. This means that the theory is inconsistent and needs to be modified at short distances or high energies. The way that string theory does this is to give up one of the basic assumptions of quantum field theory, the assumption that elementary particles are mathematical points. Instead, it is a quantum field theory of one-dimensional extended objects called strings. There are very few consistent theories of this type, but superstring theory shows great promise as a unified quantum theory of all fundamental forces including gravity. So far, nobody has constructed a realistic string theory of elementary particles that could serve as a new standard model of particles and forces, since there is much that needs to be better understood first. But that, together with a deeper understanding of cosmology, is the goal. This is very much a work in progress.

Even though string theory is not yet fully formulated, and we cannot yet give a detailed description of how the standard model of elementary particles should emerge at low energies, there are some general features of the theory that can be identified. These are features that seem to be quite generic irrespective of how various details are resolved. The first, and perhaps most important, is that general relativity is necessarily incorporated in the theory. It gets modified at very short distances/high energies but at ordinary distances and energies it is present in exactly the form proposed by Einstein. This is significant, because it is arising within the framework of a consistent quantum theory. Ordinary quantum field theory does not allow gravity to exist; string theory requires it. The second general fact is that Yang–Mills gauge theories of the sort that comprise the standard model naturally arise in string theory. We do not understand why the specific Yang–Mills gauge theory based on the symmetry group $SU(3) \times SU(2) \times U(1)$ should be preferred, but (anomaly-free) theories of this general type do arise naturally at ordinary energies. The third general feature of string theory solutions is that they possess a special kind of symmetry called supersymmetry. The mathematical consistency of string theory depends crucially on supersymmetry, and it is very hard to find consistent solutions (i.e., quantum vacua) that do not preserve at least a portion of this supersymmetry. This prediction of string theory differs from the other two (general relativity and gauge theories) in that it really is a prediction. It is a generic feature of string theory that has not yet been observed experimentally.

I. SUPERSYMMETRY

Even though supersymmetry is a very important part of the story, the discussion here will be very brief. Like the electroweak symmetry in the standard model, supersymmetry is necessarily a broken symmetry. A variety of arguments, not specific to string theory, suggest that the characteristic energy scale associated to supersymmetry breaking should be related to the electroweak scale, in other words, in the range 100 GeV–1 TeV. (Recall that the rest mass of a proton or neutron corresponds to an energy of approximately 1 GeV. Also, the masses of the $W^{\pm}$ and Z^0 particles, which transmit the weak nuclear forces, correspond to energies of approximately 100 GeV.) Supersymmetry implies that all known elementary particles should have partner particles whose masses are in this general range. If supersymmetry were not broken, these particles would have exactly the same masses as the known particles, and that is definitely excluded. This means that some of these superpartners should be observable at the CERN Large Hadron Collider (LHC), which is scheduled to begin operating in 2005 or 2006. There is even a chance that Fermilab Tevatron experiments could find superparticles before then. (CERN is a lab outside of Geneva, Switzerland and Fermilab is located outside of Chicago, IL.)

In most versions of phenomenological supersymmetry there is a multiplicatively conserved quantum number called R-parity. All known particles have even R-parity, whereas their superpartners have odd R-parity. This implies that the superparticles must be pair-produced in particle collisions. It also implies that the lightest supersymmetry particle (or LSP) should be absolutely stable. It is not known with certainty which superparticle is the LSP, but one popular guess is that it is a "neutralino."

This is an electrically neutral fermion that is a quantum-mechanical mixture of the partners of the photon, Z^0, and neutral Higgs particles. Such an LSP would interact very weakly, more or less like a neutrino. It is of considerable interest, since it has properties that make it an excellent dark matter candidate. There are experimental searches underway in Europe and in the United States for a class of dark matter particles called WIMPS (weakly interacting massive particles). Since the LSP is of an example of a WIMP, these searches could discover the LSP some day. However, the current experiments might not have sufficient detector volume to compensate for the exceedingly small LSP cross sections, so we may have to wait for future upgrades of the detectors.

There are three unrelated arguments that point to the same 100 GeV–1 TeV mass range for superparticles. The one we have just been discussing, a neutralino LSP as an important component of dark matter, requires a mass of about 100 GeV. The precise number depends on the mixture that comprises the LSP, what their density is, and a number of other details. A second argument is based on a theoretical issue called the hierarchy problem. This is the fact that in the standard model quantum corrections tend to renormalize the Higgs mass to an unacceptably high value. The way to prevent this is to extend the standard model to a supersymmetric standard model and to have the supersymmetry be broken at a scale comparable to the Higgs mass, and hence to the electroweak scale. This works because the quantum corrections to the Higgs mass are more mild in the supersymmetric version of the theory. The third argument that gives an estimate of the supersymmetry-breaking scale is based on grand unification. If one accepts the notion that the standard model gauge group is embedded in a larger group such as $SU(5)$ or $SO(10)$, which is broken at a high mass scale, then the three standard model coupling constants should unify at that mass scale. Given the spectrum of particles, one can compute the variation of the couplings as a function of energy using renormalization group equations. One finds that if one only includes the standard model particles this unification fails quite badly. However, if one also includes all the supersymmetry particles required by the minimal supersymmetric extension of the standard model, then the couplings do unify at an energy of about 2×10^{16} GeV. This is a very striking success. For this agreement to take place, it is necessary that the masses of the superparticles are less than a few TeV.

There is other support for this picture, such as the ease with which supersymmetric grand unification explains the masses of the top and bottom quarks and electroweak symmetry breaking. Despite all these indications, we cannot be certain that supersymmetry at the electroweak scale really is correct until it is demonstrated experimentally. One could suppose that all the successes that we have listed are a giant coincidence, and the correct description of TeV scale physics is based on something entirely different. The only way we can decide for sure is by doing the experiments. I am optimistic that supersymmetry will be found, and that the experimental study of the detailed properties of the superparticles will teach us a great deal.

A. Basic Ideas of String Theory

In conventional quantum field theory the elementary particles are mathematical points, whereas in perturbative string theory the fundamental objects are one-dimensional loops (of zero thickness). Strings have a characteristic length scale, which can be estimated by dimensional analysis. Since string theory is a relativistic quantum theory that includes gravity it must involve the fundamental constants c (the speed of light), $\hbar$ (Planck's constant divided by 2π), and G (Newton's gravitational constant). From these one can form a length, known as the Planck length

$$\ell_p = \left(\frac{\hbar G}{c^3}\right)^{3/2} = 1.6 \times 10^{-33}\,\text{cm}. \qquad (1)$$

Similarly, the Planck mass is

$$m_p = \left(\frac{\hbar c}{G}\right)^{1/2} = 1.2 \times 10^{19}\,\text{GeV/c}^2. \qquad (2)$$

Experiments at energies far below the Planck energy cannot resolve distances as short as the Planck length. Thus, at such energies, strings can be accurately approximated by point particles. From the viewpoint of string theory, this explains why quantum field theory has been so successful.

As a string evolves in time it sweeps out a two-dimensional surface in spacetime, which is called the world sheet of the string. This is the string counterpart of the world line for a point particle. In quantum field theory, analyzed in perturbation theory, contributions to amplitudes are associated to Feynman diagrams, which depict possible configurations of world lines. In particular, interactions correspond to junctions of world lines. Similarly, perturbative string theory involves string world sheets of various topologies. A particularly significant fact is that these world sheets are generically smooth. The existence of interaction is a consequence of world-sheet topology rather than a local singularity on the world sheet. This difference from point-particle theories has two important implications. First, in string theory the structure of interactions is uniquely determined by the free theory. There are no arbitrary interactions to be chosen. Second, the occurrence of ultraviolet divergences in point-particle quantum field theories can be traced to the fact that interactions are associated to world-line junctions at specific spacetime points. Because the string world sheet is smooth, without

any singular behavior at short distances, string theory has no ultraviolet divergences.

B. A Brief History of String Theory

String theory arose in the late 1960s out of an attempt to describe the strong nuclear force, which acts on a class of particles called hadrons. The first string theory that was constructed only contained bosons. The construction of a better string theory that also includes fermions led to the discovery of supersymmetric strings (later called superstrings) in 1971. The subject fell out of favor around 1973 with the development of quantum chromodynamics (QCD), which was quickly recognized to be the correct theory of strong interactions. Also, string theories had various peculiar features, such as extra dimensions and massless particles, which are not appropriate for a hadron theory.

Among the massless string states there is one that corresponds to a particle with two units of spin. In 1974, it was shown by Joël Scherk and the author (Scherk and Schwarz, 1974), and independently by Yoneya (1974), that this particle interacts like a graviton, so that string theory actually contains general relativity. This led us to propose that string theory should be used for unification of all elementary particles and forces rather than as a theory of hadrons and the strong nuclear force. This implied, in particular, that the string length scale should be comparable to the Planck length, rather than the size of hadrons (10^{-13} cm), as had been previously assumed.

In the period now known as the "first superstring revolution," which took place in 1984–1985, there were a number of important developments (described later in this article) that convinced a large segment of the theoretical physics community that this is a worthy area of research. By the time the dust settled in 1985 we had learned that there are five distinct consistent string theories, and that each of them requires spacetime supersymmetry in the ten dimensions (nine spatial dimensions plus time). The theories, which will be described later, are called type I, type IIA, type IIB, $SO(32)$ heterotic, and $E_8 \times E_8$ heterotic. In the "second superstring revolution," which took place around 1995, we learned that the five string theories are actually special solutions of a completely unique underlying theory.

C. Compactification

In the context of the original goal of string theory—to explain hadron physics—extra dimensions are unacceptable. However, in a theory that incorporates general relativity, the geometry of spacetime is determined dynamically. Thus one could imagine that the theory admits consistent quantum solutions in which the six extra spatial dimensions form a compact space, too small to have been observed. The natural first guess is that the size of this space should be comparable to the string scale and the Planck length. Since the equations of the theory must be satisfied, the geometry of this six-dimensional space is not arbitrary. A particularly appealing possibility, which is consistent with the equations, is that it forms a type of space called a Calabi–Yau space (Candelas *et al.*, 1985).

Calabi–Yau compactification, in the context of the $E_8 \times E_8$ heterotic string theory, can give a low-energy effective theory that closely resembles a supersymmetric extension of the standard model. There is actually a lot of freedom, because there are very many different Calabi–Yau spaces, and there are other arbitrary choices that can be made. Still, it is interesting that one can come quite close to realistic physics. It is also interesting that the number of quark and lepton families that one obtains is determined by the topology of the Calabi–Yau space. Thus, for suitable choices, one can arrange to end up with exactly three families. People were very excited by this scenario in 1985. Today, we tend to make a more sober appraisal that emphasizes all the arbitrariness that is involved, and the things that don't work exactly right. Still, it would not be surprising if some aspects of this picture survive as part of the story when we understand the right way to describe the real world.

D. Perturbation Theory

Until 1995 it was only understood how to formulate string theories in terms of perturbation expansions. Perturbation theory is useful in a quantum theory that has a small dimensionless coupling constant, such as quantum electrodynamics, since it allows one to compute physical quantities as power series expansions in the small parameter. In quantum electrodynamics (QED) the small parameter is the fine-structure constant $\alpha \sim 1/137$. Since this is quite small, perturbation theory works very well for QED. For a physical quantity $T(\alpha)$, one computes (using Feynman diagrams)

$$T(\alpha) = T_0 + \alpha T_1 + \alpha^2 T_2 + \cdots. \tag{3}$$

It is the case generically in quantum field theory that expansions of this type are divergent. More specifically, they are asymptotic expansions with zero radius convergence. Nonetheless, they can be numerically useful if the expansion parameter is small. The problem is that there are various nonperturbative contributions (such as instantons) that have the structure

$$T_{NP} \sim e^{-(\text{const.}/\alpha)}. \tag{4}$$

In a theory such as QCD, there are problems for which perturbation theory is useful (due to asymptotic freedom) and

other ones where it is not. For problems of the latter type, such as computing the hadron spectrum, nonperturbative methods of computation, such as lattice gauge theory, are required.

In the case of string theory the dimensionless string coupling constant, denoted g_s, is determined dynamically by the expectation value of a scalar field called the dilaton. There is no particular reason that this number should be small. So it is unlikely that a realistic vacuum could be analyzed accurately using perturbation theory. More importantly, these theories have many qualitative properties that are inherently nonperturbative. So one needs nonperturbative methods to understand them.

E. The Second Superstring Revolution

Around 1995 some amazing and unexpected "dualities" were discovered that provided the first glimpses into nonperturbative features of string theory. These dualities were quickly recognized to have three major implications.

The dualities enabled us to relate all five of the superstring theories to one another. This meant that, in a fundamental sense, they are all equivalent to one another. Another way of saying this is that there is a unique underlying theory, and what we had been calling five theories are better viewed as perturbation expansions of this underlying theory about five different points (in the space of consistent quantum vacua). This was a profoundly satisfying realization, since we really didn't want five theories of nature. That there is a completely unique theory, without any dimensionless parameters, is the best outcome for which one could have hoped. To avoid confusion, it should be emphasized that even though the theory is unique, it is entirely possible that there are many consistent quantum vacua. Classically, the corresponding statement is that a unique equation can admit many solutions. It is a particular solution (or quantum vacuum) that ultimately must describe nature. At least, this is how a particle physicist would say it. If we hope to understand the origin and evolution of the universe, in addition to properties of elementary particles, it would be nice if we could also understand cosmological solutions.

A second crucial discovery was that the theory admits a variety of nonperturbative excitations, called p-branes, in addition to the fundamental strings. The letter p labels the number of spatial dimensions of the excitation. Thus, in this language, a point particle is a 0-brane, a string is a 1-brane, and so forth. The reason that p-branes were not discovered in perturbation theory is that they have tension (or energy density) that diverges as $g_s \to 0$. Thus they are absent from the perturbative theory.

The third major discovery was that the underlying theory also has an eleven-dimensional solution, which is called M-theory. Later, we will explain how the eleventh dimension arises.

One type of duality is called S duality. (The choice of the letter S has no great significance.) Two string theories (let's call them A and B) are related by S duality if one of them evaluated at strong coupling is equivalent to the other one evaluated at weak coupling. Specifically, for any physical quantity f, one has

$$f_A(g_s) = f_B(1/g_s). \tag{5}$$

Two of the superstring theories—type I and $SO(32)$ heterotic—are related by S duality in this way. The type IIB theory is self-dual. Thus S duality is a symmetry of the IIB theory, and this symmetry is unbroken if $g_s = 1$. Thanks to S duality, the strong coupling behavior of each of these three theories is determined by a weak-coupling analysis. The remaining two theories, type IIA and $E_8 \times E_8$ heterotic, behave very differently at strong coupling. They grow an eleventh dimension.

Another astonishing duality, which goes by the name of T duality, was discovered several years earlier. It can be understood in perturbation theory, which is why it was found first. But, fortunately, it often continues to be valid even at strong coupling. T duality can relate different compactifications of different theories. For example, suppose theory A has a compact dimension that is a circle of radius R_A and theory B has a compact dimension that is a circle of radius R_B. If these two theories are related by T duality this means that they are equivalent provided that

$$R_A R_B = (\ell_s)^2, \tag{6}$$

where ℓ_s is the fundamental string length scale. This has the amazing implication that when one of the circles becomes small the other one becomes large. Later, we will explain how this is possible. T duality relates the two type II theories and the two heterotic theories. There are more complicated examples of the same phenomenon involving compact spaces that are more complicated than a circle, such as tori, K3, Calabi–Yau spaces, etc.

F. The Origins of Gauge Symmetry

There are a variety of mechanisms than can give rise to Yang–Mills type gauge symmetries in string theory. Here, we will focus on two basic possibilities: Kaluza–Klein symmetries and brane symmetries.

The basic Kaluza–Klein idea goes back to the 1920s, though it has been much generalized since then. The idea is to suppose that the ten- or eleven-dimensional geometry has a product structure $M \times K$, where M is Minkowski spacetime and K is a compact manifold. Then, if K has symmetries, these appear as gauge symmetries of the effective theory defined on M. The Yang–Mills gauge fields

arise as components of the gravitational metric field with one direction along K and the other along M. For example, if the space K is an n-dimensional sphere, the symmetry group is $SO(n+1)$, if it is CP^n—which has $2n$ dimensions—it is $SU(n+1)$, and so forth. Elegant as this may be, it seems unlikely that a realistic K has any such symmetries. Calabi–Yau spaces, for example, do not have any.

A rather more promising way of achieving realistic gauge symmetries is via the brane approach. Here the idea is that a certain class of p-branes (called D-branes) have gauge fields that are restricted to their world volume. This means that the gauge fields are not defined throughout the ten- or eleven-dimensional spacetime but only on the $(p+1)$-dimensional hypersurface defined by the D-branes. This picture suggests that the world we observe might be a D-brane embedded in a higher dimensional space. In such a scenario, there can be two kinds of extra dimensions: compact dimensions along the brane and compact dimensions perpendicular to the brane.

The traditional viewpoint, which in my opinion is still the best bet, is that all extra dimensions (of both types) have sizes of order 10^{-30}–10^{-32} cm corresponding to an energy scale of 10^{16}–10^{18} GeV. This makes them inaccessible to direct observation, though their existence would have definite low-energy consequences. However, one can and should ask "what are the experimental limits?" For compact dimensions along the brane, which support gauge fields, the nonobservation of extra dimensions in tests of the standard model implies a bound of about 1 TeV. The LHC should extend this to about 10 TeV. For compact dimensions "perpendicular to the brane," which only support excitations with gravitational strength forces, the best bounds come from Cavendish-type experiments, which test the $1/R^2$ structure of the Newton force law at short distances. No deviations have been observed to a distance of about 1 mm so far. Experiments planned in the near future should extend the limit to about 100 μ. Obviously, observation of any deviation from $1/R^2$ would be a major discovery.

G. Conclusion

This introductory section has sketched some of the remarkable successes that string theory has achieved over the past 30 years. There are many others that did not fit in this brief survey. Despite all this progress, there are some very important and fundamental questions whose answers are unknown. It seems that whenever a breakthrough occurs, a host of new questions arise, and the ultimate goal still seems a long way off. To convince you that there is a long way to go, let us list some of the most important questions.

- What is the theory? Even though a great deal is known about string theory and M-theory, it seems that the optimal formulation of the underlying theory has not yet been found. It might be based on principles that have not yet been formulated.
- We are convinced that supersymmetry is present at high energies and probably at the electroweak scale, too. But we do not know how or why it is broken.
- A very crucial problem concerns the energy density of the vacuum, which is a physical quantity in a gravitational theory. This is characterized by the cosmological constant, which observationally appears to have a small positive value—so that the vacuum energy of the universe is comparable to the energy in matter. In Planck units this is a tiny number ($\Lambda \sim 10^{-120}$). If supersymmetry were unbroken, we could argue that $\Lambda = 0$, but if it is broken at the 1 TeV scale, that would seem to suggest $\Lambda \sim 10^{-60}$, which is very far from the truth. Despite an enormous amount of effort and ingenuity, it is not yet clear how superstring theory will conspire to break supersymmetry at the TeV scale and still give a value for Λ that is much smaller than 10^{-60}. The fact that the desired result is about the square of this might be a useful hint.
- Even though the underlying theory is unique, there seem to be many consistent quantum vacua. We would very much like to formulate a theoretical principle (not based on observation) for choosing among these vacua. It is not known whether the right approach to the answer is cosmological, probabilistic, anthropic, or something else.

II. STRING THEORY BASICS

In this section we will describe the world-sheet dynamics of the original bosonic string theory. As we will see this theory has various unrealistic and unsatisfactory properties. Nonetheless it is a useful preliminary before describing supersymmetric strings, because it allows us to introduce many of the key concepts without simultaneously addressing the added complications associated with fermions and supersymmetry.

We will describe string dynamics from a first-quantized world-sheet sum-over-histories point of view. This approach is closely tied to perturbation theory analysis. It should be contrasted with "second quantized" string field theory, which is based on field operators that create or destroy entire strings. To explain the methodology, let us begin by reviewing the world-line description a massive point particle.

A. World-Line Description of a Massive Point Particle

A point particle sweeps out a trajectory (or world line) in spacetime. This can be described by functions $x^\mu(\tau)$ that describe how the world line, parameterized by τ, is embedded in the spacetime, whose coordinates are denoted x^μ. For simplicity, let us assume that the spacetime is flat Minkowski space with a Lorentz metric

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{7}$$

Then, the Lorentz invariant line element is given by

$$ds^2 = -\eta_{\mu\nu}\, dx^\mu\, dx^\nu. \tag{8}$$

In units $\hbar = c = 1$, the action for a particle of mass m is given by

$$S = -m \int ds. \tag{9}$$

This could be generalized to a curved spacetime by replacing $\eta_{\mu\nu}$ by a metric $g_{\mu\nu}(x)$, but we will not do so here. In terms of the embedding functions, $x^\mu(\tau)$, the action can be rewritten in the form

$$S = -m \int d\tau \sqrt{-\eta_{\mu\nu}\dot{x}^\mu \dot{x}^\nu}, \tag{10}$$

where dots represent τ derivatives. An important property of this action is invariance under local reparametrizations. This is a kind of gauge invariance, whose meaning is that the form of S is unchanged under an arbitrary reparametrization of the world line $\tau \to \tau(\tilde{\tau})$. Actually, one should require that the function $\tau(\tilde{\tau})$ is smooth and monotonic ($\frac{d\tau}{d\tilde{\tau}} > 0$). The reparametrization invariance is a one-dimensional analog of the four-dimensional general coordinate invariance of general relativity. Mathematicians refer to this kind of symmetry as diffeomorphism invariance.

The reparametrization invariance of S allows us to choose a gauge. A nice choice is the "static gauge"

$$x^0 = \tau. \tag{11}$$

In this gauge (renaming the parameter t) the action becomes

$$S = -m \int \sqrt{1 - v^2}\, dt, \tag{12}$$

where

$$\vec{v} = \frac{d\vec{x}}{dt}. \tag{13}$$

Requiring this action to be stationary under an arbitrary variation of $\vec{x}(t)$ gives the Euler–Lagrange equations

$$\frac{d\vec{p}}{dt} = 0, \tag{14}$$

where

$$\vec{p} = \frac{\delta S}{\delta \vec{v}} = \frac{m\vec{v}}{\sqrt{1 - v^2}}, \tag{15}$$

which is the usual result. So we see that standard relativistic kinematics follows from the action $S = -m \int ds$.

B. World-Volume Actions

We can now generalize the analysis of the massive point particle to a p-brane of tension T_p. The action in this case involves the invariant $(p+1)$-dimensional volume and is given by

$$S_p = -T_p \int d\mu_{p+1}, \tag{16}$$

where the invariant volume element is

$$d\mu_{p+1} = \sqrt{-\det\left(-\eta_{\mu\nu}\partial_\alpha x^\mu \partial_\beta x^\nu\right)}\, d^{p+1}\sigma. \tag{17}$$

Here the embedding of the p-brane into d-dimensional spacetime is given by functions $x^\mu(\sigma^\alpha)$. The index $\alpha = 0, \ldots, p$ labels the $p+1$ coordinates σ^α of the p-brane world-volume and the index $\mu = 0, \ldots, d-1$ labels the d coordinates x^μ of the d-dimensional spacetime. We have defined

$$\partial_\alpha x^\mu = \frac{\partial x^\mu}{\partial \sigma^\alpha}. \tag{18}$$

The determinant operation acts on the $(p+1) \times (p+1)$ matrix whose rows and columns are labeled by α and β. The tension T_p is interpreted as the mass per unit volume of the p-brane. For a 0-brane, it is just the mass. The action S_p is reparametrization invariant. In other words, substituting $\sigma^\alpha = \sigma^\alpha(\tilde{\sigma}^\beta)$, it takes the same form when expressed in terms of the coordinates $\tilde{\sigma}^\alpha$.

Let us now specialize to the string, $p = 1$. Evaluating the determinant gives

$$S[x] = -T \int d\sigma\, d\tau \sqrt{\dot{x}^2 x'^2 - (\dot{x} \cdot x')^2}, \tag{19}$$

where we have defined $\sigma^0 = \tau$, $\sigma^1 = \sigma$, and

$$\dot{x}^\mu = \frac{\partial x^\mu}{\partial \tau}, \qquad x'^\mu = \frac{\partial x^\mu}{\partial \sigma}. \tag{20}$$

This action, called the Nambu–Goto action, was first proposed in 1970 (Nambu, 1970 and Goto, 1971). The Nambu–Goto action is equivalent to the action

$$S[x, h] = -\frac{T}{2} \int d^2\sigma \sqrt{-h} h^{\alpha\beta} \eta_{\mu\nu} \partial_\alpha x^\mu \partial_\beta x^\nu, \tag{21}$$

where $h_{\alpha\beta}(\sigma,\tau)$ is the world-sheet metric, $h=\det h_{\alpha\beta}$, and $h^{\alpha\beta}$ is the inverse of $h_{\alpha\beta}$. The Euler–Lagrange equations obtained by varying $h^{\alpha\beta}$ are

$$T_{\alpha\beta}=\partial_\alpha x\cdot\partial_\beta x-\frac{1}{2}h_{\alpha\beta}h^{\gamma\delta}\partial_\gamma x\cdot\partial_\delta x=0. \tag{22}$$

The equation $T_{\alpha\beta}=0$ can be used to eliminate the world-sheet metric from the action, and when this is done one recovers the Nambu–Goto action. (To show this take the determinant of both sides of the equation $\partial_\alpha x\cdot\partial_\beta x=\frac{1}{2}h_{\alpha\beta}h^{\gamma\delta}\partial_\gamma x\cdot\partial_\delta x$.)

In addition to reparametrization invariance, the action $S[x,h]$ has another local symmetry, called conformal invariance (or Weyl invariance). Specifically, it is invariant under the replacement

$$\begin{aligned} h_{\alpha\beta}&\to\Lambda(\sigma,\tau)h_{\alpha\beta}\\ x^\mu&\to x^\mu. \end{aligned} \tag{23}$$

This local symmetry is special to the $p=1$ case (strings).

The two reparametrization invariance symmetries of $S[x,h]$ allow us to choose a gauge in which the three functions $h_{\alpha\beta}$ (this is a symmetric 2×2 matrix) are expressed in terms of just one function. A convenient choice is the "conformally flat gauge"

$$h_{\alpha\beta}=\eta_{\alpha\beta}e^{\phi(\sigma,\tau)}. \tag{24}$$

Here $\eta_{\alpha\beta}$ denotes the two-dimensional Minkowski metric of a flat world-sheet. However, because of the factor e^ϕ, $h_{\alpha\beta}$ is only "conformally flat." Classically, substitution of this gauge choice into $S[x,h]$ yields the gauge-fixed action

$$S=\frac{T}{2}\int d^2\sigma\,\eta^{\alpha\beta}\partial_\alpha x\cdot\partial_\beta x. \tag{25}$$

Quantum mechanically, the story is more subtle. Instead of eliminating h via its classical field equations, one should perform a Feynman path integral, using standard machinery to deal with the local symmetries and gauge fixing. When this is done correctly, one finds that in general ϕ does not decouple from the answer. Only for the special case $d=26$ does the quantum analysis reproduce the formula we have given based on classical reasoning (Polyakov, 1981). Otherwise, there are correction terms whose presence can be traced to a conformal anomaly (i.e., a quantum-mechanical breakdown of the conformal invariance).

The gauge-fixed action [Eq. (25)] is quadratic in the x's. Mathematically, it is the same as a theory of d free scalar fields in two dimensions. The equations of motion obtained by varying x^μ are simply free two-dimensional wave equations:

$$\ddot{x}^\mu-x''^\mu=0. \tag{26}$$

This is not the whole story, however, because we must also take account of the constraints $T_{\alpha\beta}=0$. Evaluated in the conformally flat gauge, these constraints are

$$\begin{aligned} T_{01}&=T_{10}=\dot{x}\cdot x'=0\\ T_{00}&=T_{11}=\frac{1}{2}(\dot{x}^2+x'^2)=0. \end{aligned} \tag{27}$$

Adding and subtracting gives

$$(\dot{x}\pm x')^2=0. \tag{28}$$

C. Boundary Conditions

To go further, one needs to choose boundary conditions. There are three important types. For a closed string one should impose periodicity in the spatial parameter σ. Choosing its range to be π (as is conventional)

$$x^\mu(\sigma,\tau)=x^\mu(\sigma+\pi,\tau). \tag{29}$$

For an open string (which has two ends), each end can be required to satisfy either Neumann or Dirichlet boundary conditions (for each value of μ).

$$\text{Neumann:}\quad \frac{\partial x^\mu}{\partial\sigma}=0\qquad\text{at}\quad\sigma=0\text{ or }\pi \tag{30}$$

$$\text{Dirichlet:}\quad \frac{\partial x^\mu}{\partial\tau}=0\qquad\text{at}\quad\sigma=0\text{ or }\pi. \tag{31}$$

The Dirichlet condition can be integrated, and then it specifies a spacetime location on which the string ends. The only way this makes sense is if the open string ends on a physical object—it ends on a D-brane. (D stands for Dirichlet.) If all the open-string boundary conditions are Neumann, then the ends of the string can be anywhere in the spacetime. The modern interpretation is that this means that there are spacetime-filling D-branes present.

Let us now consider the closed-string case in more detail. The general solution of the two-dimensional wave equation is given by a sum of "right-movers" and "left-movers":

$$x^\mu(\sigma,\tau)=x_R^\mu(\tau-\sigma)+x_L^\mu(\tau+\sigma). \tag{32}$$

These should be subject to the following additional conditions:

1. $x^\mu(\sigma,\tau)$ is real
2. $x^\mu(\sigma+\pi,\tau)=x^\mu(\sigma,\tau)$
3. $(x_L')^2=(x_R')^2=0$; these are the $T_{\alpha\beta}=0$ constraints in Eq. (28)

The first two conditions can be solved explicitly in terms of Fourier series:

$$x_R^\mu = \frac{1}{2}x^\mu + \ell_s^2 p^\mu(\tau-\sigma) + \frac{i}{\sqrt{2}}\ell_s \sum_{n\neq 0}\frac{1}{n}\alpha_n^\mu e^{-2in(\tau-\sigma)}$$
$$x_L^\mu = \frac{1}{2}x^\mu + \ell_s^2 p^\mu(\tau+\sigma) + \frac{i}{\sqrt{2}}\ell_s \sum_{n\neq 0}\frac{1}{n}\tilde{\alpha}_n^\mu e^{-2in(\tau+\sigma)}, \tag{33}$$

where the expansion parameters α_n^μ, $\tilde{\alpha}_n^\mu$ satisfy

$$\alpha_{-n}^\mu = \left(\alpha_n^\mu\right)^\dagger, \qquad \tilde{\alpha}_{-n}^\mu = \left(\tilde{\alpha}_n^\mu\right)^\dagger. \tag{34}$$

The center-of-mass coordinate x^μ and momentum p^μ are also real. The fundamental string length scale ℓ_s is related to the tension T by

$$T = \frac{1}{2\pi\alpha'}, \qquad \alpha' = \ell_s^2. \tag{35}$$

The parameter α' is called the universal Regge slope, since the string modes lie on linear parallel Regge trajectories with this slope.

D. Quantization

The analysis of closed-string left-moving modes, closed-string right-moving modes, and open-string modes are all very similar. Therefore, to avoid repetition, we will focus on the closed-string right-movers. Starting with the gauge-fixed action in Eq. (25), the canonical momentum of the string is

$$p^\mu(\sigma,\tau) = \frac{\delta S}{\delta \dot{x}^\mu} = T\dot{x}^\mu. \tag{36}$$

Canonical quantization (this is just free two-dimensional field theory for scalar fields) gives

$$[p^\mu(\sigma,\tau), x^\nu(\sigma',\tau)] = -i\hbar\eta^{\mu\nu}\delta(\sigma-\sigma'). \tag{37}$$

In terms of the Fourier modes (setting $\hbar = 1$) these become

$$[p^\mu, x^\nu] = -i\eta^{\mu\nu} \tag{38}$$
$$\begin{aligned}\left[\alpha_m^\mu, \alpha_n^\nu\right] &= m\delta_{m+n,0}\,\eta^{\mu\nu},\\ \left[\tilde{\alpha}_m^\mu, \tilde{\alpha}_n^\nu\right] &= m\delta_{m+n,0}\,\eta^{\mu\nu},\end{aligned} \tag{39}$$

and all other commutators vanish.

Recall that a quantum-mechanical harmonic oscillator can be described in terms of raising and lowering operators, usually called $a^\dagger$ and a, which satisfy

$$[a, a^\dagger] = 1. \tag{40}$$

We see that, aside from a normalization factor, the expansion coefficients α_{-m}^μ and α_m^μ are raising and lowering operators. There is just one problem. Because $\eta^{00} = -1$, the time components are proportional to oscillators with the wrong sign ($[a, a^\dagger] = -1$). This is potentially very bad, because such oscillators create states of negative norm, which could lead to an inconsistent quantum theory (with negative probabilities, etc.). Fortunately, as we will explain, the $T_{\alpha\beta} = 0$ constraints eliminate the negative-norm states from the physical spectrum.

The classical constraint for the right-moving closed-string modes, $(x_R')^2 = 0$, has Fourier components

$$L_m = \frac{T}{2}\int_0^\pi e^{-2im\sigma}(x_R')^2\,d\sigma = \frac{1}{2}\sum_{n=-\infty}^{\infty}\alpha_{m-n}\cdot\alpha_n, \tag{41}$$

which are called Virasoro operators. Since α_m^μ does not commute with α_{-m}^μ, L_0 needs to be normal-ordered:

$$L_0 = \frac{1}{2}\alpha_0^2 + \sum_{n=1}^{\infty}\alpha_{-n}\cdot\alpha_n. \tag{42}$$

Here $\alpha_0^\mu = \ell_s p^\mu/\sqrt{2}$, where p^μ is the momentum.

E. The Free String Spectrum

Recall that the Hilbert space of a harmonic oscillator is spanned by states $|n\rangle$, $n = 0, 1, 2, \ldots$, where the ground state, $|0\rangle$, is annihilated by the lowering operator ($a\,|\,0\rangle = 0$) and

$$|n\rangle = \frac{(a^\dagger)^n}{\sqrt{n!}}\,|\,0\rangle. \tag{43}$$

Then, for a normalized ground-state ($\langle 0\,|\,0\rangle = 1$), one can use $[a, a^\dagger] = 1$ repeatedly to prove that

$$\langle m\,|\,n\rangle = \delta_{m,n} \tag{44}$$

and

$$a^\dagger a\,|\,n\rangle = n\,|\,n\rangle. \tag{45}$$

The string spectrum (of right-movers) is given by the product of an infinite number of harmonic-oscillator Fock spaces, one for each α_n^μ, subject to the Virasoro constraints (Virasoro, 1970)

$$\begin{aligned}(L_0 - q)\,|\,\phi\rangle &= 0\\ L_n\,|\,\phi\rangle &= 0, \quad n > 0.\end{aligned} \tag{46}$$

Here $|\phi\rangle$ denotes a physical state, and q is a constant to be determined. It accounts for the arbitrariness in the normal-ordering prescription used to define L_0. As we will see, the L_0 equation is a generalization of the Klein–Gordon equation. It contains $p^2 = -\partial\cdot\partial$ plus oscillator terms whose eigenvalue will determine the mass of the state.

It is interesting to work out the algebra of the Virasoro operators L_m, which follows from the oscillator algebra. The result, called the Virasoro algebra, is

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}(m^3 - m)\delta_{m+n,0}. \quad (47)$$

The second term on the right-hand side is called the "conformal anomaly term" and the constant c is called the "central charge" or "conformal anomaly." Each component of x^μ contributes one unit to the central charge, so that altogether $c = d$.

There is a more sophisticated way to describe the string spectrum (in terms of BRST cohomology), but it is equivalent to the more elementary approach presented here. In the BRST approach, gauge-fixing to the conformal gauge in the quantum theory requires the addition of world-sheet Faddeev-Popov ghosts, which turn out to contribute $c = -26$. Thus the total conformal anomaly of the x^μ and the ghosts cancels for the particular choice $d = 26$, as we asserted earlier. Moreover, it is also necessary to set the parameter $q = 1$, so that mass-shell condition becomes

$$(L_0 - 1)\,|\,\phi\rangle = 0. \quad (48)$$

Since the mathematics of the open-string spectrum is the same as that of closed-string right-movers, let us now use the equations we have obtained to study the open-string spectrum. (Here we are assuming that the open-string boundary conditions are all Neumann, corresponding to spacetime-filling D-branes.) The mass-shell condition is

$$M^2 = -p^2 = -\frac{1}{2}\alpha_0^2 = N - 1, \quad (49)$$

where

$$N = \sum_{n=1}^{\infty} \alpha_{-n} \cdot \alpha_n = \sum_{n=1}^{\infty} n a_n^\dagger \cdot a_n. \quad (50)$$

The $a^\dagger$'s and a's are properly normalized raising and lowering operators. Since each $a^\dagger a$ has eigenvalues 0, 1, 2, ..., the possible values of N are also 0, 1, 2, ... The unique way to realize $N = 0$ is for all the oscillators to be in the ground state, which we denote simply by $|0; p^\mu\rangle$, where p^μ is the momentum of the state. This state has $M^2 = -1$, which is a tachyon (p^μ is spacelike). Such a faster-than-light particle is certainly not possible in a consistent quantum theory, because the vacuum would be unstable. However, in perturbation theory (which is the framework we are implicitly considering) this instability is not visible. Since this string theory is only supposed to be a warm-up exercise before considering tachyon-free superstring theories, let us continue without worrying about the vacuum instability.

The first excited state, with $N = 1$, corresponds to $M^2 = 0$. The only way to achieve $N = 1$ is to excite the first oscillator once:

$$|\phi\rangle = \zeta_\mu \alpha_{-1}^\mu\,|\,0; p\rangle. \quad (51)$$

Here ζ_μ denotes the polarization vector of a massless spin-one particle. The Virasoro constraint condition $L_1\,|\,\phi\rangle = 0$ implies that ζ_μ must satisfy

$$p^\mu \zeta_\mu = 0. \quad (52)$$

This ensures that the spin is transversely polarized, so there are $d-2$ independent polarization states. This agrees with what one finds for a massless Maxwell or Yang–Mills field.

At the next mass level, where $N = 2$ and $M^2 = 1$, the most general possibility has the form

$$|\phi\rangle = \left(\zeta_\mu \alpha_{-2}^\mu + \lambda_{\mu\nu}\alpha_{-1}^\mu \alpha_{-1}^\nu\right)|\,0; p\rangle. \quad (53)$$

However, the constraints $L_1\,|\,\phi\rangle = L_2\,|\,\phi\rangle = 0$ restrict ζ_μ and $\lambda_{\mu\nu}$. The analysis is interesting, but only the results will be described. If $d > 26$, the physical spectrum contains a negative-norm state, which is not allowed. However, when $d = 26$, this state becomes zero-norm and decouples from the theory. This leaves a pure massive "spin two" (symmetric traceless tensor) particle as the only physical state at this mass level.

Let us now turn to the closed-string spectrum. A closed-string state is described as a tensor product of a left-moving state and a right-moving state, subject to the condition that the N value of the left-moving and the right-moving state is the same. The reason for this "level-matching" condition is that we have $(L_0 - 1)\,|\,\phi\rangle = (\tilde{L}_0 - 1)\,|\,\phi\rangle = 0$. The sum $(L_0 + \tilde{L}_0 - 2)\,|\,\phi\rangle$ is interpreted as the mass-shell condition, while the difference $(L_0 - \tilde{L}_0)|\phi\rangle = (N - \tilde{N})\,|\,\phi\rangle = 0$ is the level-matching condition.

Using this rule, the closed-string ground state is just

$$|0\rangle \otimes |\,0\rangle, \quad (54)$$

which represents a spin 0 tachyon with $M^2 = -2$. (The notation no longer displays the momentum p of the state.) Again, this signals an unstable vacuum, but we will not worry about it here. Much more important, and more significant, is the first excited state

$$|\phi\rangle = \zeta_{\mu\nu}\left(\alpha_{-1}^\mu\,|\,0\rangle \otimes \tilde{\alpha}_{-1}^\nu\,|\,0\rangle\right), \quad (55)$$

which has $M^2 = 0$. The Virasoro constraints $L_1\,|\,\phi\rangle = \tilde{L}_1\,|\,\phi\rangle = 0$ imply that $p^\mu \zeta_{\mu\nu} = 0$. Such a polarization tensor encodes three distinct spin states, each of which plays a fundamental role in string theory. The symmetric part of $\zeta_{\mu\nu}$ encodes a spacetime metric field $g_{\mu\nu}$ (massless spin two) and a scalar dilaton field ϕ (massless spin zero). The

$g_{\mu\nu}$ field is the graviton field, and its presence (with the correct gauge invariances) accounts for the fact that the theory contains general relativity, which is a good approximation for energies well below the string scale. Its vacuum value determines the spacetime geometry. Similarly, the value of ϕ determines the string coupling constant ($g_s = \langle e^{\phi} \rangle$).

$\zeta_{\mu\nu}$ also has an antisymmetric part, which corresponds to a massless antisymmetric tensor gauge field $B_{\mu\nu} = -B_{\nu\mu}$. This field has a gauge transformation of the form

$$\delta B_{\mu\nu} = \partial_\mu \Lambda_\nu - \partial_\nu \Lambda_\mu, \tag{56}$$

(which can be regarded as a generalization of the gauge transformation rule for the Maxwell field: $\delta A_\mu = \partial_\mu \Lambda$). The gauge-invariant field strength (analogous to $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$) is

$$H_{\mu\nu\rho} = \partial_\mu B_{\nu\rho} + \partial_\nu B_{\rho\mu} + \partial_\rho B_{\mu\nu}. \tag{57}$$

The importance of the $B_{\mu\nu}$ field resides in the fact that the fundamental string is a source for $B_{\mu\nu}$, just as a charged particle is a source for the vector potential A_μ. Mathematically, this is expressed by the coupling

$$q \int B_{\mu\nu}\, dx^\mu \wedge dx^\nu, \tag{58}$$

which generalizes the coupling of a charged particle to a Maxwell field

$$q \int A_\mu\, dx^\mu. \tag{59}$$

F. The Number of Physical States

The number of physical states grows rapidly as a function of mass. This can be analyzed quantitatively. For the open string, let us denote the number of physical states with $\alpha' M^2 = n - 1$ by d_n. These numbers are encoded in the generating function

$$G(w) = \sum_{n=0}^{\infty} d_n w^n = \prod_{m=1}^{\infty} (1 - w^m)^{-24}. \tag{60}$$

The exponent 24 reflects the fact that in 26 dimensions, once the Virasoro conditions are taken into account, the spectrum is exactly what one would get from 24 transversely polarized oscillators. It is easy to deduce from this generating function the asymptotic number of states for large n, as a function of n

$$d_n \sim n^{-27/4} e^{4\pi\sqrt{n}}. \tag{61}$$

This asymptotic degeneracy implies that the finite-temperature partition function

$$\mathrm{tr}\,(e^{-\beta H}) = \sum_{n=0}^{\infty} d_n e^{-\beta M_n} \tag{62}$$

diverges for $\beta^{-1} = T > T_H$, where T_H is the Hagedorn temperature

$$T_H = \frac{1}{4\pi\sqrt{\alpha'}} = \frac{1}{4\pi \ell_s}. \tag{63}$$

T_H might be the maximum possible temperature or else a critical temperature at which there is a phase transition.

G. The Structure of String Perturbation Theory

As we discussed in the first section, perturbation theory calculations are carried out by computing Feynman diagrams. Whereas in ordinary quantum field theory Feynman diagrams are webs of world lines, in the case of string theory they are two-dimensional surfaces representing string world-sheets. For these purposes, it is convenient to require that the world-sheet geometry is Euclidean (i.e., the world-sheet metric $h_{\alpha\beta}$ is positive definite). The diagrams are classified by their topology, which is very well understood in the case of two-dimensional surfaces. The world-sheet topology is characterized by the number of handles (h), the number of boundaries (b), and whether or not they are orientable. The order of the expansion (i.e., the power of the string-coupling constant) is determined by the Euler number of the world sheet M. It is given by $\chi(M) = 2 - 2h - b$. For example, a sphere has $h = b = 0$, and hence $\chi = 2$. A torus has $h = 1$, $b = 0$, and $\chi = 0$, a cylinder has $h = 0$, $b = 2$, and $\chi = 0$, and so forth. Surfaces with $\chi = 0$ admit a flat metric.

A scattering amplitude is given by a path integral of the schematic structure

$$\int Dh_{\alpha\beta}(\sigma) Dx^\mu(\sigma) e^{-S[h,x]} \prod_{i=1}^{n_c} \int_M V_{\alpha_i}(\sigma_i)\, d^2\sigma_i \prod_{j=1}^{n_o} \times \int_{\partial M} V_{\beta_j}(\sigma_j)\, d\sigma_j. \tag{64}$$

The action $S[h, x]$ is given in Eq. (21). V_{α_i} is a vertex operator that describes emission or absorption of a closed-string state of type α_i from the interior of the string world-sheet, and V_{β_j} is a vertex operator that describes emission of absorption of an open-string state of type β_j from the boundary of the string world-sheet. There are lots of technical details that are not explained here. In the end, one finds that the conformally inequivalent world-sheets of a given topology are described by a finite number of parameters, and thus these amplitudes can be recast as finite-dimensional integrals over these "moduli." (The momentum integrals are already done.) The dimension of the resulting integral turns out to be

$$N = 3(2h + b - 2) + 2n_c + n_o. \tag{65}$$

As an example consider the amplitude describing elastic scattering of two open-string ground states. In this case

$h=0$, $b=1$, $n_c=0$, $n_o=4$, and therefore $N=1$. In terms of the usual Mandelstam invariants $s=-(p_1+p_2)^2$ and $t=-(p_1-p_4)^2$, the result is

$$A(s,t)=g_s^2\int_0^1 dx x^{-\alpha(s)-1}(1-x)^{-\alpha(t)-1}, \qquad (66)$$

where the Regge trajectory $\alpha(s)$ is

$$\alpha(s)=1+\alpha' s. \qquad (67)$$

This integral is just the Euler beta function

$$A(s,t)=g_s^2 B(-\alpha(s),-\alpha(t))=g_s^2\frac{\Gamma(-\alpha(s))\Gamma(-\alpha(t))}{\Gamma(-\alpha(s)-\alpha(t))}. \qquad (68)$$

This is the famous Veneziano amplitude (Veneziano, 1968), which got the whole subject started.

H. Recapitulation

This section described some of the basic facts of the 26-dimensional bosonic string theory. One significant point that has not yet been made clear is that there are actually a number of distinct theories depending on what kinds of strings one includes.

- Oriented closed strings only
- Oriented closed-strings and oriented open-strings; in this case one can incorporate $U(n)$ gauge symmetry
- Unoriented closed strings only
- Unoriented closed-strings and unoriented open-strings; in this case one can incorporate $SO(n)$ or $Sp(n)$ gauge symmetry

As we have mentioned already, all the bosonic string theories are unphysical as they stand, because (in each case) the closed-string spectrum contains a tachyon. A tachyon means that one is doing perturbation theory about an unstable vacuum. This is analogous to the unbroken symmetry extremum of the Higgs potential in the standard model. In that case, we know that there is a stable minimum, where the Higgs fields acquires a vacuum value. Recently, there has been success in demonstrating that open-string tachyons condense at a stable minimum, but the fate of the closed-string tachyon is still an open problem.

III. SUPERSTRINGS

Among the deficiencies of the bosonic string theory is the fact that there are no fermions. As we will see, the addition of fermions leads quite naturally to supersymmetry and hence superstrings. There are two alternative formalisms that are used to study superstrings. The original one, which grew out of the 1971 papers by Ramond and by Neveu and Schwarz (1971) is called the RNS formalism. In this approach, the supersymmetry of the two-dimensional world-sheet theory plays a central role. The second approach, developed by Michael Green and the author in the early 1980s (Green and Schwarz, 1981), emphasizes supersymmetry in the ten-dimensional spacetime. Which one is more useful depends on the problem being studied. Only the RNS approach will be presented here.

In the RNS formalism, the world-sheet theory is based on the d functions $x^\mu(\sigma,\tau)$ that describe the embedding of the world-sheet in the spacetime, just as before. However, in order to supersymmetrize the world-sheet theory, we also introduce d fermionic partner fields $\psi^\mu(\sigma,\tau)$. Note that x^μ transforms as a vector from the spacetime viewpoint, but as d scalar fields from the two-dimensional world-sheet viewpoint. The ψ^μ also transform as a spacetime vector, but as world-sheet spinors. Altogether, x^μ and ψ^μ described d supersymmetry multiplets, one for each value of μ.

The reparametrization invariant world-sheet action described in the preceding section can be generalized to have local supersymmetry on the world-sheet, as well. (The details of how that works are a bit too involved to describe here.) When one chooses a suitable conformal gauge ($h_{\alpha\beta}=e^\phi\eta_{\alpha\beta}$), together with an appropriate fermionic gauge condition, one ends up with a world-sheet theory that has global supersymmetry supplemented by constraints. The constraints form a super-Virasoro algebra. This means that in addition to the Virasoro constraints of the bosonic string theory, there are fermionic constraints, as well.

A. The Gauge-Fixed Theory

The globally supersymmetric world-sheet action that arises in the conformal gauge takes the form

$$S=-\frac{T}{2}\int d^2\sigma\left(\partial_\alpha x^\mu\partial^\alpha x_\mu-i\bar\psi^\mu\rho^\alpha\partial_\alpha\psi_\mu\right). \qquad (69)$$

The first term is exactly the same as in Eq. (25) of the bosonic string theory. Recall that it has the structure of d free scalar fields. The second term that has now been added is just d free massless spinor fields, with Dirac-type actions. The notation is that ρ^α are two 2×2 Dirac matrices and $\psi=\binom{\psi_-}{\psi_+}$ is a two-component Majorana spinor. The Majorana condition simply means that ψ_+ and ψ_- are real in a suitable representation of the Dirac algebra. In fact, a convenient choice is one for which

$$\bar\psi\rho^\alpha\partial_\alpha\psi=\psi_-\partial_+\psi_-+\psi_+\partial_-\psi_+, \qquad (70)$$

where $\partial_\pm$ represent derivatives with respect to $\sigma^\pm = \tau \pm \sigma$. In this basis, the equations of motion are simply

$$\partial_+ \psi_-^\mu = \partial_- \psi_+^\mu = 0. \tag{71}$$

Thus ψ_-^μ describes right-movers and ψ_+^μ describes left-movers.

Concentrating on the right-movers ψ_-^μ, the global supersymmetry transformations, which are a symmetry of the gauge-fixed action, are

$$\begin{aligned} \delta x^\mu &= i\epsilon \psi_-^\mu \\ \delta \psi_-^\mu &= -2\partial_- x^\mu \epsilon. \end{aligned} \tag{72}$$

It is easy to show that this is a symmetry of the action [Eq. (69)]. There is an analogous symmetry for the left-movers. (Accordingly, the world-sheet theory is said to have (1, 1) supersymmetry.) Continuing to focus on the right-movers, the Virasoro constraint is

$$(\partial_- x)^2 + \frac{i}{2} \psi_-^\mu \partial_- \psi_{\mu -} = 0. \tag{73}$$

The first term is what we found in the bosonic string theory, and the second term is an additional fermionic contribution. There is also an associated fermionic constraint

$$\psi_-^\mu \partial_- x_\mu = 0. \tag{74}$$

The Fourier modes of these constraints generate the super-Virasoro algebra. There is a second identical super-Virasoro algebra for the left-movers.

As in the bosonic string theory, the Virasoro algebra has conformal anomaly terms proportional to a central charge c. As in that theory, each component of x^μ contributes $+1$ to the central charge, for a total of d, while (in the BRST quantization approach) the reparametrization symmetry ghosts contribute -26. But now there are additional contributions. Each component of ψ^μ gives $+1/2$, for a total of $d/2$, and the local supersymmetry ghosts contribute $+11$. Adding all of this up, gives a grand total of $c = \frac{3d}{2} - 15$. Thus, we see that the conformal anomaly cancels for the specific choice $d = 10$. This is the preferred critical dimension for superstrings, just as $d = 26$ is the critical dimension for bosonic strings. For other values the theory has a variety of inconsistencies.

B. The R and NS Sectors

Let us now consider boundary conditions for $\psi^\mu(\sigma, \tau)$. (The story for x^μ is exactly as before.) First, let us consider open-string boundary conditions. For the action to be well-defined, it turns out that one must set $\psi_+ = \pm \psi_-$ at the two ends $\sigma = 0, \pi$. An overall sign is a matter of convention, so we can set

$$\psi_+^\mu(0, \tau) = \psi_-^\mu(0, \tau), \tag{75}$$

without loss of generality. But this still leaves two possibilities for the other end, which are called R and NS:

$$\begin{aligned} \text{R: } & \psi_+^\mu(\pi, \tau) = \psi_-^\mu(\pi, \tau) \\ \text{NS: } & \psi_+^\mu(\pi, \tau) = -\psi_-^\mu(\pi, \tau). \end{aligned} \tag{76}$$

Combining these with the equations of motion $\partial_- \psi_+ = \partial_+ \psi_- = 0$, allows us to express the general solutions as Fourier series

$$\begin{aligned} \text{R: } \psi_-^\mu &= \frac{1}{\sqrt{2}} \sum_{n \in \mathbf{Z}} d_n^\mu e^{-in(\tau - \sigma)} \\ \psi_+^\mu &= \frac{1}{\sqrt{2}} \sum_{n \in \mathbf{Z}} d_n^\mu e^{-in(\tau + \sigma)} \\ \text{NS: } \psi_-^\mu &= \frac{1}{\sqrt{2}} \sum_{r \in \mathbf{Z} + 1/2} b_r^\mu e^{-ir(\tau - \sigma)} \\ \psi_+^\mu &= \frac{1}{\sqrt{2}} \sum_{r \in \mathbf{Z} + 1/2} b_r^\mu e^{-ir(\tau + \sigma)}. \end{aligned} \tag{77}$$

The Majorana condition implies that $d_{-n}^\mu = d_n^{\mu\dagger}$ and $b_{-r}^\mu = b_r^{\mu\dagger}$. Note that the index n takes integer values, whereas the index r takes half-integer values $(\pm\frac{1}{2}, \pm\frac{3}{2}, \ldots)$. In particular, only the R boundary condition gives a zero mode.

Canonical quantization of the free fermi fields $\psi^\mu(\sigma, \tau)$ is very standard and straightforward. The result can be expressed as anticommutation relations for the coefficients d_m^μ and b_r^μ:

$$\begin{aligned} \text{R: } & \left\{ d_n^\mu, d_n^\nu \right\} = \eta^{\mu\nu} \delta_{m+n,0} \qquad m, n \in \mathbf{Z} \\ \text{NS: } & \left\{ d_r^\mu, d_s^\nu \right\} = \eta^{\mu\nu} \delta_{r+s,0} \qquad r, s \in \mathbf{Z} + \frac{1}{2}. \end{aligned} \tag{78}$$

Thus, in addition to the harmonic oscillator operators α_m^μ that appear as coefficients in mode expansions of x^μ, there are fermionic oscillator operators d_m^μ or b_r^μ that appear as coefficients in mode expansions of ψ^μ. The basic structure $\{b, b^\dagger\} = 1$ is very simple. It describes a two-state system with $b \mid 0\rangle = 0$, and $b^\dagger \mid 0\rangle = \mid 1\rangle$. The b's or d's with negative indices can be regarded as raising operators and those with positive indices as lowering operators, just as we did for the α_n^μ.

In the NS sector, the ground state $|0; p\rangle$ satisfies

$$\alpha_m^\mu \mid 0; p\rangle = b_r^\mu \mid 0; p\rangle = 0, \qquad m, r > 0 \tag{79}$$

which is a straightforward generalization of how we defined the ground state in the bosonic string theory. All the excited states obtained by acting with the α and b raising operators are spacetime bosons. We will see later that the ground state, defined as we have done here, is again a tachyon. However, in this theory, as we will also see, there is a way by which this tachyon can (and must) be removed from the physical spectrum.

In the R sector there are zero modes that satisfy the algebra

$$\{d_0^\mu, d_0^\nu\} = \eta^{\mu\nu}. \tag{80}$$

This is the d-dimensional spacetime Dirac algebra. Thus the d_0's should be regarded as Dirac matrices and all states in the R sector should be spinors in order to furnish representation spaces on which these operators can act. The conclusion, therefore, is that whereas all string states in the NS sector are spacetime bosons, all string states in the R sector are spacetime fermions.

In the closed-string case, the physical states are obtained by tensoring right- and left-movers, each of which are mathematically very similar to the open-string spectrum. This means that there are four distinct sectors of closed-string states: NS $\otimes$ NS and R $\otimes$ R describe spacetime bosons, whereas NS $\otimes$ R and R $\otimes$ NS describe spacetime fermions. We will return to explore what this gives later, but first we need to explore the right-movers by themselves in more detail.

The zero mode of the fermionic constraint $\psi^\mu \partial_- x_\mu = 0$ gives a wave equation for (fermionic) strings in the Ramond sector, $F_0|\psi\rangle = 0$, which is called the Dirac–Ramond equation. In terms of the oscillators

$$F_0 = \alpha_0 \cdot d_0 + \sum_{n\neq 0} \alpha_{-n} \cdot d_n. \tag{81}$$

The zero-mode piece of F_0, $\alpha_0 \cdot d_0$, has been isolated, because it is just the usual Dirac operator, $\gamma^\mu \partial_\mu$, up to normalization. (Recall that α_0^μ is proportional to $p_\mu = -i\partial_\mu$, and d_0^μ is proportional to the Dirac matrices γ^μ.) The fermionic ground state $|\psi_0\rangle$, which satisfies

$$\alpha_n^\mu \,|\, \psi_0\rangle = d_n^\mu \,|\, \psi_0\rangle = 0, \qquad n > 0, \tag{82}$$

satisfies the wave equation

$$\alpha_0 \cdot d_0 \,|\, \psi_0\rangle = 0, \tag{83}$$

which is precisely the massless Dirac equation. Hence the fermionic ground state is a massless spinor.

C. The GSO Projection

In the NS (bosonic) sector the mass formula is

$$M^2 = N - \frac{1}{2}, \tag{84}$$

which is to be compared with the formula $M^2 = N - 1$ of the bosonic string theory. This time the number operator N has contributions from the b oscillators as well as the α oscillators. (The reason that the normal-ordering constant is $-1/2$ instead of -1 works as follows. Each transverse α oscillator contributes $-1/24$ and each transverse b oscillator contributes $-1/48$. The result follows since the bosonic theory has 24 transverse directions and the superstring theory has 8 transverse directions.) Thus the ground state, which has $N = 0$, is now a tachyon with $M^2 = -1/2$.

This is where things stood until the 1976 work of Gliozzi, Scherk, and Olive. They noted that the spectrum admits a consistent truncation (called the GSO projection), which is necessary for the consistency of the interacting theory. In the NS sector, the GSO projection keeps states with an odd number of b-oscillator excitations and removes states with an even number of b-oscillator excitation. Once this rule is implemented the only possible values of N are half integers, and the spectrum of allowed masses are integral

$$M^2 = 0, 1, 2, \ldots. \tag{85}$$

In particular, the bosonic ground state is now massless. The spectrum no longer contains a tachyon. The GSO projection also acts on the R sector, where there is an analogous restriction on the d oscillators. This amounts to imposing a chirality projection on the spinors.

Let us look at the massless spectrum of the GSO-projected theory. The ground-state boson is now a massless vector, represented by the state $\zeta_\mu b^\mu_{-1/2}\,|\,0; p\rangle$, which (as before) has $d - 2 = 8$ physical polarizations. The ground-state fermion is a massless Majorana–Weyl fermion which has $\frac{1}{4}\cdot 2^{d/2} = 8$ physical polarizations. Thus there are an equal number of bosons and fermions, as is required for a theory with spacetime supersymmetry. In fact, this is the pair of fields that enter into ten-dimensional super Yang–Mills theory. The claim is that the complete theory now has spacetime supersymmetry.

If there is spacetime supersymmetry, then there should be an equal number of bosons and fermions at every mass level. Let us denote the number of bosonic states with $M^2 = n$ by $d_{NS}(n)$ and the number of fermionic states with $M^2 = n$ by $d_R(n)$. Then we can encode these numbers in generating functions, just as we did for the bosonic string theory

$$f_{NS}(w) = \sum_{n=0}^{\infty} d_{NS}(n) w^n = \frac{1}{2\sqrt{w}}\left(\prod_{m=1}^{\infty}\left(\frac{1+w^{m-1/2}}{1-w^m}\right)^8 - \prod_{m=1}^{\infty}\left(\frac{1-w^{m-1/2}}{1-w^m}\right)^8\right) \tag{86}$$

$$f_R(w) = \sum_{n=0}^{\infty} d_R(n) w^n = 8\prod_{m=1}^{\infty}\left(\frac{1+w^m}{1-w^m}\right)^8. \tag{87}$$

The 8's in the exponents refer to the number of transverse directions in ten dimensions. The effect of the GSO projection is the subtraction of the second term in f_{NS} and the

reduction of the coefficient in f_R from 16 to 8. In 1829, Jacobi discovered the formula

$$f_R(w) = f_{NS}(w). \tag{88}$$

(He used a different notation, of course.) For him this relation was an obscure curiosity, but we now see that it tells us that the number of bosons and fermions is the same at every mass level, which provides strong evidence for supersymmetry of this string theory in ten dimensions. A complete proof of supersymmetry for the interacting theory was constructed by Green and the author five years after the GSO paper (Green and Schwarz, 1981).

D. Type II Superstrings

We have described the spectrum of bosonic (NS) and fermionic (R) string states. This also gives the spectrum of left- and right-moving closed-string modes, so we can form the closed-string spectrum by forming tensor products as before. In particular, the massless right-moving spectrum consists of a vector and a Majorana–Weyl spinor. Thus the massless closed-string spectrum is given by

$$(\text{vector} + \text{MW spinor}) \otimes (\text{vector} + \text{MW spinor}). \tag{89}$$

There are actually two distinct possibilities, because the two MW spinors can have either opposite chirality or the same chirality.

When the two MW spinors have opposite chirality, the theory is called type IIA superstring theory, and its massless spectrum forms the type IIA supergravity multiplet. This theory is left-right symmetric. In other words, the spectrum is invariant under mirror reflection. This implies that the IIA theory is parity conserving. When the two MW spinors have the same chirality, the resulting type IIB superstring theory is chiral, and hence parity violating. In each case there are two gravitinos, arising from vector $\otimes$ spinor and spinor $\otimes$ vector, which are gauge fields for local supersymmetry. Thus, since both type II superstring theories have two gravitinos, they have local $\mathcal{N} = 2$ supersymmetry in the ten-dimensional sense. The supersymmetry charges are Majorana–Weyl spinors, which have 16 real components, so the type II theories have 32 conserved supercharges. This is the same amount of supersymmetry as what is usually called $\mathcal{N} = 8$ in four dimensions, and it is believed to be the most that is possible in a consistent interacting theory.

The type II superstring theories contain only oriented closed strings (in the absence of D-branes). However, there is another superstring theory, called type I, which can be obtained by a projection of the type IIB theory, that only keeps the diagonal sum of the two gravitinos. Thus, this theory only has $\mathcal{N} = 1$ supersymmetry (16 supercharges). It is a theory of unoriented closed strings. However, it can be supplemented by unoriented open strings. This introduces a Yang–Mills gauge group, which classically can be $SO(n)$ or $Sp(n)$ for any value of n. Quantum consistency singles out $SO(32)$ as the unique possibility. This restriction can be understood in a number of ways. The way that it was first discovered was by considering anomalies.

E. Anomalies

Chiral (parity-violating) gauge theories can be inconsistent due to anomalies. This happens when there is a quantum mechanical breakdown of the gauge symmetry, which is induced by certain one-loop Feynman diagrams. (Sometimes one also considers breaking of global symmetries by anomalies, which does not imply an inconsistency. That is not what we are interested in here.) In the case of four dimensions, the relevant diagrams are triangles, with the chiral fields going around the loop and three gauge fields attached as external lines. In the case of the standard model, the quarks and leptons are chiral and contribute to a variety of possible anomalies. Fortunately, the standard model has just the right particle content so that all of the gauge anomalies cancel. If one omits the quark or lepton contributions, it does not work.

In the case of ten-dimensional chiral gauge theories, the potentially anomalous Feynman diagrams are hexagons, with six external gauge fields. The anomalies can be attributed to the massless fields, and therefore they can be analyzed in the low-energy effective field theory. There are several possible cases in ten dimensions:

- $\mathcal{N} = 1$ supersymmetric Yang–Mills theory. This theory has anomalies for every choice of gauge group.
- Type I supergravity. This theory has gravitational anomalies.
- Type IIA supergravity. This theory is nonchiral, and therefore it is trivially anomaly-free.
- Type IIB supergravity. This theory has three chiral fields each of which contributes to several kinds of gravitational anomalies. However, when their contributions are combined, the anomalies all cancel. (This result was obtained by Alvarez–Gaumé and Witten, 1983.)
- Type I supergravity coupled to super Yang–Mills. This theory has both gauge and gravitational anomalies for every choice of Yang–Mills gauge group except $SO(32)$ and $E_8 \times E_8$. For these two choices, all the anomalies cancel. (This result was obtained by Green and Schwarz, 1984a.)

As we mentioned earlier, at the classical level one can define type I superstring theory for any orthogonal or symplectic gauge group. Now we see that at the quantum level,

the only choice that is consistent is $SO(32)$. For any other choice there are fatal anomalies. The term $SO(32)$ is used here somewhat imprecisely. There are several different Lie groups that have the same Lie algebra. It turns out that the particular Lie group that is appropriate is Spin $(32)/\mathbf{Z}_2$. It contains one spinor conjugacy class in addition to the adjoint conjugacy class.

F. Heterotic Strings

The two Lie groups that are singled out—$E_8 \times E_8$ and Spin $(32)/\mathbf{Z}_2$—have several properties in common. Each of them has dimension $= 496$ and rank $= 16$. Moreover, their weight lattices correspond to the only two even self-dual lattices in 16 dimensions. This last fact was the crucial clue that led Gross, Harvey, Martinec, and Rohm (1985) to the discovery of the heterotic string soon after the anomaly cancellation result. One hint is the relation $10 + 16 = 26$. The construction of the heterotic string uses the $d = 26$ bosonic string for the left-movers and the $d = 10$ superstring for the right movers. The 16 extra left-moving dimensions are associated to an even self-dual 16-dimensional lattice. In this way one builds in the $SO(32)$ or $E_8 \times E_8$ gauge symmetry.

Thus, to recapitulate, by 1985 we had five consistent superstring theories, type I [with gauge group $SO(32)$], the two type II theories, and the two heterotic theories. Each is a supersymmetric ten-dimensional theory. The perturbation theory was studied in considerable detail, and while some details may not have been completed, it was clear that each of the five theories has a well-defined, ultraviolet-finite perturbation expansion, satisfying all the usual consistency requirements (unitarity, analyticity, causality, etc.). This was pleasing, though it was somewhat mysterious why there should be five consistent quantum gravity theories. It took another ten years until we understood that these are actually five special quantum vacua of a unique underlying theory.

G. T Duality

T duality, an amazing result obtained in the late 1980s, relates one string theory with a circular compact dimension of radius R to another string theory with a circular dimension of radius $1/R$ (in units $\ell_s = 1$). This is very profound, because it indicates a limitation of our usual motions of classical geometry. Strings see geometry differently from point particles. Let us examine how this is possible.

The key to understanding T duality is to consider the kinds of excitations that a string can have in the presence of a circular dimension. One class of excitations, called Kaluza–Klein excitations, is a very general feature of any quantum theory, whether or not based on strings. The idea is that in order for the wave function e^{ipx} to be single valued, the momentum along the circle must be a multiple of $1/R$, $p = n/R$, where n is an integer. From the lower dimension viewpoint this is interpreted as a contribution $(n/R)^2$ to the square of the mass.

There is a second type of excitation that is special to closed strings. Namely, a closed string can wind m times around the circular dimension, getting caught up on the topology of the space, contributing an energy given by the string tension times the length of the string

$$E_m = 2\pi R \cdot m \cdot T. \tag{90}$$

Putting $T = \frac{1}{2\pi}$ (for $\ell_s = 1$), this is just $E_m = mR$.

The combined energy-squared of the Kaluza–Klein and winding-mode excitations is

$$E^2 = \left(\frac{n}{R}\right)^2 + (mR)^2 + \cdots, \tag{91}$$

where the dots represent string oscillator contributions. Under T duality

$$m \leftrightarrow n, \qquad R \leftrightarrow 1/R. \tag{92}$$

Together, these interchanges leave the energy invariant. This means that what is interpreted as a Kaluza–Klein excitation in one string theory is interpreted as a winding-mode excitation in the T-dual theory, and the two theories have radii R and $1/R$, respectively. The two principle examples of T-dual pairs are the two type II theories and the two heterotic theories. In the latter case there are additional technicalities that explain how the two gauge groups are related. Basically, when the compactification on a circle to nine dimensions is carried out in each case, it is necessary to include effects that we haven't explained (called Wilson lines) to break the gauge groups to $SO(16) \times SO(16)$, which is a common subgroup of $SO(32)$ and $E_8 \times E_8$.

IV. FROM SUPERSTRINGS TO M-THEORY

Superstring theory is currently undergoing a period of rapid development in which important advances in understanding are being achieved. The focus in this section will be on explaining why there can be an eleven-dimensional vacuum, even though there are only ten dimensions in perturbative superstring theory. The nonperturbative extension of superstring theory that allows for an eleventh dimension has been named *M-theory*. The letter M is intended to be flexible in its interpretation. It could stand for *magic*, *mystery*, or *meta* to reflect our current state of incomplete understanding. Those who think that two-dimensional supermembranes (the M2-brane) are fundamental may regard M as standing for *membrane*. An approach called *Matrix theory* is another possibility. And, of course, some view M-theory as the *mother* of all theories.

In the first superstring revolution we identified five distinct superstring theories, each in ten dimensions. Three of them, the type I theory and the two heterotic theories, have $\mathcal{N}=1$ supersymmetry in the ten-dimensional sense. Since the minimal ten-dimensional spinor is simultaneously Majorana and Weyl, this corresponds to 16 conserved supercharges. The other two theories, called type IIA and type IIB, have $\mathcal{N}=2$ supersymmetry (32 supercharges). In the IIA case the two spinors have opposite handedness so that the spectrum is left-right symmetric (nonchiral). In the IIB case the two spinors have the same handedness and the spectrum is chiral.

In each of these five superstring theories it became clear, and was largely proved, that there are consistent perturbation expansions of on-shell scattering amplitudes. In four of the five cases (heterotic and type II) the fundamental strings are oriented and unbreakable. As a result, these theories have particularly simple perturbation expansions. Specifically, there is a unique Feynman diagram at each order of the loop expansion. The Feynman diagrams depict string world-sheets, and therefore they are two-dimensional surfaces. For these four theories the unique L-loop diagram is a closed orientable genus-L Riemann surface, which can be visualized as a sphere with L handles. External (incoming or outgoing) particles are represented by N points (or "punctures") on the Riemann surface. A given diagram represents a well-defined integral of dimension $6L+2N-6$. This integral has no ultraviolet divergences, even though the spectrum contains states of arbitrarily high spin (including a massless graviton). From the viewpoint of point-particle contributions, string and supersymmetry properties are responsible for incredible cancellations. Type I superstrings are unoriented and breakable. As a result, the perturbation expansion is more complicated for this theory, and various world-sheet diagrams at a given order have to be combined properly to cancel divergences and anomalies.

As we explained in the previous section, T duality relates two string theories when one spatial dimension forms a circle (denoted S^1). Then the ten-dimensional geometry is $R^9 \times S^1$. T duality identifies this string compactification with one of a second string theory also on $R^9 \times S^1$. If the radii of the circles in the two cases are denoted R_1 and R_2, then

$$R_1 R_2 = \alpha'. \tag{93}$$

Here $\alpha' = \ell_s^2$ is the universal Regge slope parameter, and ℓ_s is the fundamental string length scale (for both string theories). Note that T duality implies that shrinking the circle to zero in one theory corresponds to decompactification of the dual theory.

The type IIA and IIB theories are T dual, so compactifying the nonchiral IIA theory on a circle of radius R and letting $R \to 0$ gives the chiral IIB theory in ten dimensions. This means, in particular, that they should not be regarded as distinct theories. The radius R is actually the vacuum value of a scalar field, which arises as an internal component of the ten-dimensional metric tensor. Thus the type IIA and type IIB theories in ten dimensions are two limiting points in a continuous moduli space of quantum vacua. The two heterotic theories are also T dual, though (as we mentioned earlier) there are additional technical details in this case. T duality applied to the type I theory gives a dual description, which is sometimes called type I′ or IA.

A. M-Theory

In the 1970s and 1980s various supersymmetry and supergravity theories were constructed. In particular, supersymmetry representation theory showed that the largest possible spacetime dimension for a supergravity theory (with spins ≤ 2) is eleven. Eleven-dimensional supergravity, which has 32 conserved supercharges, was constructed in 1978 by Cremmer, Julia, and Scherk (1978). It has three kinds of fields—the graviton field (with 44 polarizations), the gravitino field (with 128 polarizations), and a three-index gauge field $C_{\mu\nu\rho}$ (with 84 polarizations). These massless particles are referred to collectively as the *supergraviton*. Eleven dimension supergravity is nonrenormalizable, and thus it cannot be a fundamental theory. However, we now believe that it is a low-energy effective description of M-theory, which is a well-defined quantum theory. This means, in particular, that higher dimension terms in the effective action for the supergravity fields have uniquely determined coefficients within the M-theory setting, even though they are formally infinite (and hence undetermined) within the supergravity context.

Intriguing connections between type IIA string theory and eleven dimension supergravity have been known for a long time, but the precise relationship was only explained in 1995. The field equations of eleven dimension supergravity admit a solution that describes a supermembrane. In other words, this solution has the property that the energy density is concentrated on a two-dimensional surface. A three-dimensional world-volume description of the dynamics of this supermembrane, quite analogous to the two-dimensional world volume actions of superstrings [in the GS formalism (Green and Schwarz, 1984b)], was constructed by Bergshoeff, Sezgin, and Townsend (1987) The authors suggested that a consistent eleven dimension quantum theory might be defined in terms of this membrane, in analogy to string theories in ten dimensions. (Most experts now believe that M-theory cannot be defined as a supermembrane theory.) Another striking result was that a suitable dimensional reduction of this

supermembrane gives the (previously known) type IIA superstring world-volume action. For many years these facts remained unexplained curiosities until they were reconsidered by Townsend (1995) and by Witten (1995). The conclusion is that type IIA superstring theory really does have a circular eleventh dimension in addition to the previously known ten spacetime dimensions. This fact was not recognized earlier because the appearance of the eleventh dimension is a nonperturbative phenomenon, not visible in perturbation theory.

To explain the relation between M-theory and type IIA string theory, a good approach is to identify the parameters that characterize each of them and to explain how they are related. Eleven-dimensional supergravity (and hence M-theory, too) has no dimensionless parameters. The only parameter is the eleven-dimensional Newton constant, which raised to a suitable power $(-1/9)$, gives the eleven-dimensional Planck mass m_p. When M-theory is compactified on a circle (so that the spacetime geometry is $R^{10} \times S^1$) another parameter is the radius R of the circle. Now consider the parameters of type IIA superstring theory. They are the string mass scale m_s, introduced earlier, and the dimensionless string coupling constant g_s.

We can identify compactified M-theory with type IIA superstring theory by making the following correspondences:

$$m_s^2 = 2\pi R m_p^3 \tag{94}$$

$$g_s = 2\pi R m_s. \tag{95}$$

Using these one can derive $g_s = (2\pi R m_p)^{3/2}$ and $m_s = g_s^{1/3} m_p$. The latter implies that the eleven-dimensional Planck length is shorter than the string length scale at weak coupling by a factor of $(g_s)^{1/3}$.

Conventional string perturbation theory is an expansion in powers of g_s at fixed m_s. Equation (95) shows that this is equivalent to an expansion about $R = 0$. In particular, the strong coupling limit of type IIA superstring theory corresponds to decompactification of the eleventh dimension, so in a sense M-theory is type IIA string theory at infinite coupling.* This explains why the eleventh dimension was not discovered in studies of string perturbation theory.

These relations encode some interesting facts. For one thing, the fundamental IIA string actually *is* an M2-brane of M-theory with one of its dimensions wrapped around the circular spatial dimension. Denoting the string and membrane tensions (energy per unit volume) by T_{F1} and T_{M2}, one deduces that

$$T_{F1} = 2\pi R T_{M2}. \tag{96}$$

*The $E_8 \times E_8$ heterotic string theory is also eleven-dimensional at strong coupling.

However, $T_{F1} = 2\pi m_s^2$ and $T_{M2} = 2\pi m_p^3$. Combining these relations gives Eq. (94).

B. Type II *p*-branes

Type II superstring theories contain a variety of p-brane solutions that preserve half of the 32 supersymmetries. These are solutions in which the energy is concentrated on a p-dimensional spatial hypersurface. (The world volume has $p+1$ dimensions.) The corresponding solutions of supergravity theories were constructed by Horowitz and Strominger (1991). A large class of these p-brane excitations are called *D-branes* (or *Dp*-branes when we want to specify the dimension), whose tensions are given by

$$T_{Dp} = 2\pi m_s^{p+1} / g_s. \tag{97}$$

This dependence on the coupling constant is one of the characteristic features of a D-brane. Another characteristic feature of D-branes is that they carry a charge that couples to a gauge field in the RR sector of the theory (Polchinski, 1995). The particular RR gauge fields that occur imply that p takes even values in the IIA theory and odd values in the IIB theory.

In particular, the D2-brane of the type IIA theory corresponds to the supermembrane of M-theory, but now in a background geometry in which one of the transverse dimensions is a circle. The tensions check, because [using Eqs. (94) and (95)]

$$T_{D2} = 2\pi m_s^3 / g_s = 2\pi m_p^3 = T_{M2}. \tag{98}$$

The mass of the first Kaluza–Klein excitation of the eleven-dimensional supergraviton is $1/R$. Using Eq. (95), we see that this can be identified with the D0-brane. More identifications of this type arise when we consider the magnetic dual of the M-theory supermembrane, which is a five-brane, called the M5-brane.* Its tension is $T_{M5} = 2\pi m_p^6$. Wrapping one of its dimensions around the circle gives the D4-brane, with tension

$$T_{D4} = 2\pi R T_{M5} = 2\pi m_s^5 / g_s. \tag{99}$$

If, on the other hand, the M5-frame is not wrapped around the circle, one obtains the NS5-brane of the IIA theory with tension

$$T_{NS5} = T_{M5} = 2\pi m_s^6 / g_s^2. \tag{100}$$

To summarize, type IIA superstring theory is M-theory compactified on a circle of radius $R = g_s \ell_s$. M-theory is believed to be a well-defined quantum theory in eleven-dimension, which is approximated at low energy by eleven-dimensional supergravity. Its excitations are the

*In general, the magnetic dual of a p-brane in d dimensions is a $(d-p-4)$-brane.

massless supergraviton, the M2-brane, and the M5-brane. These account both for the (perturbative) fundamental string of the IIA theory and for many of its nonperturbative excitations. The identities that we have presented here are exact, because they are protected by supersymmetry.

C. Type IIB Superstring Theory

Type IIB superstring theory, which is the other maximally supersymmetric string theory with 32 conserved supercharges, is also ten-dimensional, but unlike the IIA theory its two supercharges have the same handedness. At low-energy, type IIB superstring theory is approximated by type IIB supergravity, just as eleven-dimensional supergravity approximates M-theory. In each case the supergravity theory is only well-defined as a classical field theory, but still it can teach us a lot. For example, it can be used to construct p-brane solutions and compute their tensions. Even though such solutions are only approximate, supersymmetry considerations ensure that the tensions, which are related to the kinds of conserved charges the p-branes carry, are exact. Since the IIB spectrum contains massless chiral fields, one should check whether there are anomalies that break the gauge invariances—general coordinate invariance, local Lorentz invariance, and local supersymmetry. In fact, the UV finiteness of the string theory Feynman diagrams ensures that all anomalies must cancel, as was verified from a field theory viewpoint by Alvarez-Gaumé and Witten (1983).

Type IIB superstring theory or supergravity contains two scalar fields, the dilation ϕ and an axion χ, which are conveniently combined in a complex field

$$\rho = \chi + ie^{-\phi}. \tag{101}$$

The supergravity approximation has an $SL(2, R)$ symmetry that transforms this field nonlinearly:

$$\rho \to \frac{a\rho + b}{c\rho + d}, \tag{102}$$

where a, b, c, d are real numbers satisfying $ad - bc = 1$. However, in the quantum string theory this symmetry is broken to the discrete subgroup $SL(2, Z)$ (Hull and Townsend, 1995), which means that a, b, c, d are restricted to be integers. Defining the vacuum value of the ρ field to be

$$\langle\rho\rangle = \frac{\theta}{2\pi} + \frac{i}{g_s}, \tag{103}$$

the $SL(2, Z)$ symmetry transformation $\rho \to \rho + 1$ implies that θ is an angular coordinate. Moreover, in the special case $\theta = 0$, the symmetry transformation $\rho \to -1/\rho$ takes $g_s \to 1/g_s$. This symmetry, called *S duality*, implies that coupling constant g_s is equivalent to coupling constant $1/g_s$, so that, in the case of type II superstring theory, the weak coupling expansion and the strong coupling expansion are identical. (An analogous S-duality transformation relates the type I superstring theory to the $SO(32)$ heterotic string theory.)

Recall that the type IIA and type IIB superstring theories are T dual, meaning that if they are compactified on circles of radii R_A and R_B one obtains equivalent theories for the identification $R_A R_B = \ell_s^2$. Moreover, we saw that the type IIA theory is actually M-theory compactified on a circle. The latter fact encodes nonperturbative information. It turns out to be very useful to combine these two facts and to consider the duality between M-theory compactified on a torus ($R^9 \times T^2$) and type IIB superstring theory compactified on a circle ($R^9 \times S^1$).

A torus can be described as the complex plane modded out by the equivalence relations $z \sim z + w_1$ and $z \sim z + w_2$. Up to conformal equivalence, the periods w_1 and w_2 can be replaced by 1 and τ, with Im $\tau > 0$. In this characterization τ and $\tau' = (a\tau + b)/(c\tau + d)$, where a, b, c, d are integers satisfying $ad - bc = 1$, describe equivalent tori. Thus a torus is characterized by a modular parameter τ and an $SL(2, Z)$ modular group. The natural, and correct, conjecture at this point is that one should identify the modular parameter τ of the M-theory torus with the parameter ρ that characterizes the type IIB vacuum (Schwarz, 1995 and Aspinwall, 1996). Then the duality of M-theory and type IIB superstring theory gives a geometrical explanation of the nonperturbative S-duality symmetry of the IIB theory: the transformation $\rho \to -1/\rho$, which sends $g_s \to 1/g_s$ in the IIB theory, corresponds to interchanging the two cycles of the torus in the M theory description. To complete the story, we should relate the area of the M theory torus (A_M) to the radius of the IIB theory circle (R_B). This is a simple consequence of formulas given above

$$m_p^3 A_M = (2\pi R_B)^{-1}. \tag{104}$$

Thus the limit $R_B \to 0$, at fixed ρ, corresponds to decompactification of the M-theory torus, while preserving its shape. Conversely, the limit $A_M \to 0$ corresponds to decompactification of the IIB theory circle. The duality can be explored further by matching the various p-branes in nine-dimensions that can be obtained from either the M-theory or the IIB theory viewpoints. When this is done, one finds that everything matches nicely and that one deduces various relations among tensions (Schwarz, 1996).

Another interesting fact about the IIB theory is that it contains an infinite family of strings labeled by a pair of integers (p, q) with no common divisor (Schwarz, 1995). The $(1, 0)$ string can be identified as the fundamental IIB string, while the $(0, 1)$ string is the D-string. From this viewpoint, a (p, q) string can be regarded as a bound state

of p fundamental strings and q D-strings (Witten, 1996). These strings have a very simple interpretation in the dual M-theory description. They correspond to an M2-brane with one of its cycles wrapped around a (p, q) cycle of the torus. The minimal length of such a cycle is proportional to $|p + q\tau|$, and thus (using $\tau = \rho$) one finds that the tension of a (p, q) string is given by

$$T_{p,q} = 2\pi |p + q\rho| m_s^2. \tag{105}$$

Imagine that you lived in the nine-dimensional world that is described equivalently as M-theory compactified on a torus or as the type IIB superstring theory compactified on a circle. Suppose, moreover, you had very high energy accelerators with which you were going to determine the "true" dimension of spacetime. Would you conclude that ten or eleven is the correct answer? If either A_M or R_B was very large in Planck units there would be a natural choice, of course. But how could you decide otherwise? The answer is that either viewpoint is equally valid. What determines which choice you make is which of the massless fields you regard as "internal" components of the metric tensor and which ones you regards as matter fields. Fields that are metric components in one description correspond to matter fields in the dual one.

D. The D3-Brane and $\mathcal{N} = 4$ Gauge Theory

D-branes have a number of special properties, which make them especially interesting. By definition, they are branes on which strings can end—D stands for *Dirichlet* boundary conditions. The end of a string carries a charge, and the D-brane world-volume theory contains a $U(1)$ gauge field that carries the associated flux. When n Dp-branes are coincident, or parallel and nearly coincident, the associated $(p + 1)$-dimensional world-volume theory is a $U(n)$ gauge theory (Witten, 1996). The n^2 gauge bosons A_μ^{ij} and their supersymmetry partners arise as the ground states of oriented strings running from the ith Dp-brane to the jth Dp-brane. The diagonal elements, belonging to the Cartan subalgebra, are massless. The field A_μ^{ij} with $i \neq j$ has a mass proportional to the separation of the ith and jth branes.

The $U(n)$ gauge theory associated with a stack of n Dp-branes has maximal supersymmetry (16 supercharges). The low-energy effective theory, when the brane separations are small compared to the string scale, is supersymmetric Yang–Mills theory. These theories can be constructed by dimensional reduction of ten-dimensional supersymmetric $U(n)$ gauge theory to $p + 1$ dimensions. A case of particular interest, which we shall now focus on, is $p = 3$. A stack of n D3-branes in type IIB superstring theory has a decoupled $\mathcal{N} = 4$, $d = 4$ $U(n)$ gauge theory associated to it. This gauge theory has a number of special features. For one thing, due to boson–fermion cancellations, there are no UV divergences at any order of perturbation theory. The beta function $\beta(g)$ is identically zero, which implies that the theory is scale invariant. In fact, $\mathcal{N} = 4$, $d = 4$ gauge theories are conformally invariant. The conformal invariance combines with the supersymmetry to give a superconformal symmetry, which contains 32 fermionic generators. Another important property of $\mathcal{N} = 4$, $d = 4$ gauge theories is an electric-magnetic duality, which extends to an $SL(2, Z)$ group of dualities. Now consider the $\mathcal{N} = 4$ $U(n)$ gauge theory associated to a stack of n D3-branes in type IIB superstring theory. There is an obvious identification that turns out to be correct. Namely, the $SL(2, Z)$ duality of the gauge theory is induced from that of the ambient type IIB superstring theory. The D3-branes themselves are invariant under $SL(2, Z)$ transformations.

As we have said, a fundamental $(1, 0)$ string can end on a D3-brane. But by applying a suitable $SL(2, Z)$ transformation, this configuration is transformed to one in which a (p, q) string ends on the D3-brane. The charge on the end of this string describes a dyon with electric charge p and magnetic charge q, with respect to the appropriate gauge field. More generally, for a stack of n D3-branes, any pair can be connected by a (p, q) string. The mass is proportional to the length of the string times its tension, which we saw is proportional to $|p + q\rho|$. In this way one sees that the electrically charged particles, described by fundamental fields, belong to infinite $SL(2, Z)$ multiplets. The other states are nonperturbative excitations of the gauge theory. The field configurations that describe them preserve half of the supersymmetry. As a result their masses are given exactly by the considerations described above. An interesting question, whose answer was unknown until recently, is whether $\mathcal{N} = 4$ gauge theories in four dimensions also admit nonperturbative excitations that preserve $1/4$ of the supersymmetry. The answer turns out to be that they do, but only if $n \geq 3$. This result has a nice dual description in terms of three-string junctions (Bergman, 1998).

E. Conclusion

In this section we have described some of the interesting advances in understanding superstring theory that have taken place in the past few years. The emphasis has been on the nonperturbative appearance of an eleventh dimension in type-IIA superstring theory, as well as its implications when combined with superstring T dualities. In particular, we argued that there should be a consistent quantum vacuum, whose low-energy effective description is given by eleven-dimensional supergravity.

What we have described makes a convincing self-consistent picture, but it does not constitute a complete formulation of M-theory. In the past several years there have been some major advances in that direction, which we will

briefly mention here. The first, which goes by the name of *Matrix Theory*, bases a formulation of M-theory in flat eleven-dimensional spacetime in terms of the supersymmetric quantum mechanics of N D0-branes in the large N limit (Banks *et al.*, 1997). Matrix Theory has passed all tests that have been carried out, some of which are very nontrivial. The construction has a nice generalization to describe compactification of M-theory on a torus T^n. However, it does not seem to be useful for $n > 5$, and other compactification manifolds are (at best) awkward to handle. Another shortcoming of this approach is that it treats the eleventh dimension differently from the other ones.

Another proposal relating superstring and M-theory backgrounds to large N limits of certain field theories has been put forward by Maldacena (1997) and made more precise by Gubser, Klebanov, and Polyakov (1998), and by Witten (1998). [For a review of this subject, see (Aharony *et al.*, 2000).] In this approach, there is a conjectured duality (i.e., equivalence) between a conformally invariant field theory (CFT) in d dimensions and type IIB superstring theory or M-theory on an Anti-de-Sitter space (AdS) in $d+1$ dimensions. The remaining $9-d$ or $10-d$ dimensions form a compact space, the simplest cases being spheres. Three examples with unbroken supersymmetry are $AdS_5 \times S^5$, $AdS_4 \times S^7$, and $AdS_7 \times S^4$. This approach is sometimes referred to as *AdS/CFT duality*. This is an extremely active and very promising subject. It has already taught us a great deal about the large N behavior of various gauge theories. As usual, the easiest theories to study are ones with a lot of supersymmetry, but it appears that in this approach supersymmetry breaking is more accessible than in previous ones. For example, it might someday be possible to construct the QCD string in terms of a dual AdS gravity theory, and use it to carry out numerical calculations of the hadron spectrum. Indeed, there have already been some preliminary steps in this direction.

To sum up, I would say that despite all of the successes that have been achieved in advancing our understanding of superstring theory and M-theory, there clearly is still a long way to go. In particular, despite much effort and several imaginative proposals, we still do not have a convincing mechanism for ensuring the vanishing (or extreme smallness) of the cosmological constant for nonsupersymmetric vacua. Superstring theory is a field with very ambitious goals. The remarkable fact is that they still seem to be realistic. However, it may take a few more revolutions before they are attained.

ACKNOWLEDGMENTS

This article is based on lectures presented at the NATO Advanced Study Institute *Techniques and Concepts of High Energy Physics*, which took place in St. Croix, Virgin Islands during June 2000. The author's research is supported in part by the U.S. Dept. of Energy under Grant No. DE-FG03-92-ER40701.

SEE ALSO THE FOLLOWING ARTICLES

FIELD THEORY AND THE STANDARD MODEL • GROUP THEORY, APPLIED • PERTURBATION THEORY • QUANTUM THEORY • RELATIVITY, GENERAL

BIBLIOGRAPHY

Aharony, O., Gubser, S. S., Maldacena, J., Ooguri, H., and Oz, Y. (2000). *Phys. Rep.* **323,** 183.
Aspinwall, P. S. (1996). *Nucl. Phys. Proc. Suppl.* **46,** 30, hep-th/9508154.
Alvarez-Gaumé, L., and Witten, E. (1983). *Nucl. Phys.* **B234,** 269.
Banks, T., Fischler, W., Shenker, S., and Susskind, L. (1997). *Phys. Rev.* **D55,** 5112, hep-th/9610043.
Bergman, O. (1998). *Nucl. Phys.* **B525,** 104, hep-th/9712211.
Bergshoeff, E., Sezgin, E., and Townsend, P. K. (1987). *Phys. Lett.* **B189,** 75.
Candelas, P., Horowitz, G. T., Strominger, A., and Witten, E. (1985). *Nucl. Phys.* **B258,** 46.
Cremmer, E., Julia, B., and Scherk, J. (1978). *Phys. Lett.* **76B,** 409.
Gliozzi, F., Scherk, J., and Olive, D. (1976). *Phys. Lett.* **65B,** 282.
Goto, T. (1971). *Prog. Theor. Phys.* **46,** 1560.
Green, M. B., and Schwarz, J. H. (1984a). *Phys. Lett.* **149B,** 117.
Green, M. B., and Schwarz, J. H. (1984b). *Phys. Lett.* **136B,** 367.
Green, M. B., and Schwarz, J. H. (1981). *Nucl. Phys.* **B181,** 502; *Nucl. Phys.* **B198,** (1982) 252; *Phys. Lett.* **109B,** 444.
Green, M. B., Schwarz, J. H., and Witten, E. (1987). "Superstring Theory," in 2 vols., Cambridge Univ. Press, U.K.
Gross, D. J., Harvey, J. A., Martinec, E., and Rohm, R. (1985). *Phys. Rev. Lett.* **54,** 502.
Gubser, S. S., Klebanov, I. R., and Polyakov, A. M. (1998). *Phys. Lett.* **B428,** 105, hep-th/9802109.
Horowitz, G. T., and Strominger, A. (1991). *Nucl. Phys.* **B360,** 197.
Hull, C., and Townsend, P. (1995). *Nucl. Phys.* **B438,** 109, hep-th/9410167.
Maldacena, J. (1998). *Adv. Theor. Phys.* **2,** 231, hep-th/9711200.
Nambu, Y. (1970). Notes prepared for the Copenhagen High Energy Symposium.
Neveu, A., and Schwarz, J. H. (1971). *Nucl. Phys.* **B31,** 86.
Polchinski, J. (1995). *Phys. Rev. Lett.* **75,** 4724, hep-th/9510017.
Polchinski, J. (1998). "String Theory," in 2 vols., Cambridge Univ. Press, U.K.
Polyakov, A. M. (1981). *Phys. Lett.* **103B,** 207.
Ramond, P. (1971). *Phys. Rev.* **D3,** 2415.
Scherk, J., and Schwarz, J. H. (1974). *Nucl. Phys.* **B81,** 118.
Schwarz, J. H. (1995). *Phys. Lett.* **B360,** 13, Erratum: *Phys. Lett.* **B364,** 252, hep-th/9508143.
Schwarz, J. H. (1996). *Phys. Lett.* **B367,** 97, hep-th/9510086.
Townsend, P. K. (1995). *Phys. Lett.* **B350,** 184, hep-th/9501068.
Virasoro, M. (1970). *Phys. Rev.* **D1,** 2933.
Veneziano, G. (1968). *Nuovo Cim.* **57A,** 190.
Witten, E. (1995). *Nucl. Phys.* **B443,** 85, hep-th/9503124.
Witten, E. (1996). *Nucl. Phys.* **B460,** 335, hep-th/9510135.
Witten, E. (1998). *Adv. Theor. Math. Phys.* **2,** 253, hep-th/9802150.
Yoneya, T. (1974). *Prog. Theor. Phys.* **51,** 1907.



Thermodynamics

Stanley I. Sandler
University of Delaware

I. Thermodynamic Systems and Properties
II. Mass and Energy Flows and the Equilibrium State
III. Laws of Thermodynamics
IV. Criteria for Equilibrium and Stability
V. Pure Component Properties
VI. Phase Equilibrium in One-Component Systems
VII. Thermodynamics of Mixtures and Phase Equilibrium
VIII. Mixture Phase Equilibrium Calculations
IX. Chemical Equilibrium
X. Electrolyte Solutions
XI. Coupled Reactions

GLOSSARY

Activity coefficient A measure of the extent to which the fugacity of a species in a mixture departs from ideal mixture or ideal Henry's law behavior.
Equilibrium state A state in which there is no measurable change of properties and no flows.
Excess property The difference between the property in a mixture and that for an ideal mixture at the same temperature, pressure, and composition.
Homogeneous system A system of uniform properties.
Ideal mixture A mixture in which there is no change in volume, internal energy, or enthalpy of forming a mixture from its pure components at constant pressure at all temperatures and compositions.
Intensive property (or state variable) A property of a system that is independent of the mass of the system.
Multiphase system A heterogeneous system consisting of several phases, each of which is homogeneous.
Partial molar property The amount by which an extensive property of the system increases on the addition of an infinitesimal amount of a substance at constant temperature and pressure, expressed on a molar basis.

CHEMICAL THERMODYNAMICS is a science that is both simple and elegant and can be used to describe a large variety of physical and chemical phenomena at or near equilibrium. The basis of thermodynamics is a small set of laws based on experimental observation. These general

laws combined with constitutive relations—that is, relations that describe how properties (for example, the density) of a substance depend on the state of the system such as its temperature and pressure—allow scientists and engineers to calculate the work and heat flows accompanying a change of state and to identify the equilibrium state.

I. THERMODYNAMIC SYSTEMS AND PROPERTIES

Thermodynamics is the study of changes that occur in some part of the universe we designate as the system; everything else is the surroundings. A real or imagined boundary may separate the system from its surroundings. A collection of properties such as temperature, pressure, composition, density, refractive index, and other properties to be discussed later characterize the thermodynamic state of a system. The state of aggregation of the system (that is, whether it is a gas, liquid or solid) is referred to as its *phase*. A system may be composed of more than one phase, in which case it is a heterogeneous system; a homogeneous system consists of only a single phase. Of most interest in thermodynamics are the changes that occur with a change in temperature, state of aggregation, composition (due to chemical reaction), and/or energy of the system.

Any element of matter contains three types of energy. First is its kinetic energy, which depends on its velocity and is given by $\frac{1}{2}mv^2$, where m is the mass and v is its center-of-mass velocity (though there may be an additional contribution due to rotational motion that we will not consider). A second contribution is the potential energy, denoted by $m\phi$ and due to gravity or electric and magnetic fields. The third, and generally the most important in thermodynamics, is the internal energy U (or internal energy per unit mass $\hat{U}$), which depends on the temperature, state of aggregation, and chemical composition of the substance. In thermodynamics, one is interested in changes in internal energy between two states of the system. For changes of state that do not involve chemical reaction, a reference state of zero internal energy can be chosen arbitrarily. However, if chemical reactions do occur, the reference state for the calculation of internal energies and other properties of each substance in the reaction must be chosen in such a way that the calculated changes on reaction equal the measured values.

There are many mechanisms by which the properties of a system can change. The mass of a system can change if mass flows into or out of the system across the system boundaries. Concentrations can change as a result of mass flows, volume changes, or chemical reaction. The energy of a system can change as a result of a number of different processes. As mass flows across the system boundary, each element of mass carries its properties, such as its internal and kinetic energy. Heat (thermal energy) can cross the system boundary by direct contact (conduction and convection) or by radiation. Work or mechanical energy can be done on a system by compressing the system boundaries, by a drive shaft that crosses the system boundaries (as in a turbine or motor), or can be added as electrical energy (in a battery or electrochemical cell). Or a system can do work on its surroundings by any of these mechanisms.

A system that does not exchange mass with its surroundings is said to be closed. A system that does not exchange thermal energy with its surroundings is referred to as an *adiabatic system*. A system that is of constant volume, adiabatic, and closed is called an *isolated system*. A system whose properties are the same throughout is referred to as a *uniform system*.

It is useful to distinguish between two types of system properties. Temperature, pressure, refractive index, and density are examples of intensive properties—properties that do not depend on the size or extent of the system. Mass, volume, and total internal energy are examples of extensive properties—properties that depend on the total size of the system. Extensive properties can be converted to intensive properties by dividing by the total mass or number of moles in the system. Volume per unit mass (reciprocal of density) and internal energy per mole are examples of intensive properties. Intensive properties are also known as *state variables*. Intensive variables per unit mass will be denoted with a $\hat{}$ (as in $\hat{V}$, to denote volume per unit mass), while those on a per mole basis are given an underbar (as in $\underline{U}$, to denote internal energy per mole). Also, $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ will be used to indicate state properties such as $\underline{U}$ and $\underline{V}$, and T and P. A characteristic of a state property that is central to thermodynamic analyses is that its numerical value depends only on the state, not on the path used to get to that state. Consequently, in computing the change in value of a state property between two states, any convenient path between those states may be used, instead of the actual path.

An important experimental observation is that the specification of two independent state properties of a closed, uniform, one-component system completely fixes the values of the other state properties. For example, if two systems of the same substance in the same state of aggregation are at the same temperature and at the same pressure, all other state properties of the two systems, such as density, volume per unit mass, refractive index, internal energy per unit mass, and other properties that will be introduced shortly, will also be identical. To fix the size of the system,

one must also specify the value of one extensive variable (i.e., total mass, total volume, etc.).

II. MASS AND ENERGY FLOWS AND THE EQUILIBRIUM STATE

Flows into or out of a system can be of two types. One is a forced flow, as when a pump or other device creates a continual mechanical, thermal, or chemical driving force that results in a flow of mass or energy across the boundary of a system. The other type of flow, which we refer to as a *natural flow*, occurs into or out of a system as a result of an initial difference of some property between the system and its surroundings that in time will dissipate as a result of the flow. For example, if two metal blocks of different temperatures are put in contact, a flow of heat will occur from the block of higher temperature to the one of lower temperature until an equilibrium state is reached in which both blocks have the same temperature.

An important observation is that a closed isolated system, if initially nonuniform, will eventually reach a time-invariant state that is uniform (homogeneous system) or composed of several phases, each of which is of uniform properties. Such a state of time-invariant uniformity is the equilibrium state. Systems open to natural flows will also, in time, come to equilibrium. However, a system subjected to a continuous forced flow may in time come to a time-invariant, nonuniform steady state.

The methods of thermodynamics are used to identify, describe, and sometimes predict equilibrium states. These same methods can also be used to describe nonequilibrium and steady states provided that at each point in space and time the same relations between the state properties exist as they do in equilibrium. This implies that the internal relaxation times in the fluid must be fast compared to the time scales for changes imposed upon the system.

III. LAWS OF THERMODYNAMICS

There are four laws or experimental observations on which thermodynamics is based, though they are not always referred to as such. The first observation is that in all transformations, or changes of state, total mass is conserved (note that this need not be true in nuclear reactions, but these will not be considered here.) The second observation, the first law of thermodynamics, is that in all transformations (again, except nuclear reactions) total energy is conserved. This has been known since the experiments of J. M. Joule over the period from 1837 to 1847.

The next observation, which leads to the second law, is that all systems not subject to forced flows or imposed gradients (of temperature, pressure, concentration, velocity, etc.) will eventually evolve to a state of thermodynamic equilibrium. Also, systems in stable equilibrium states will not spontaneously change into a nonequilibrium state. For example, an isolated block of metal with a temperature gradient will evolve to a state of uniform temperature, but not vice versa. The third law of thermodynamics is of a different character than the first two and is mentioned later.

A. Mass Balance

After choosing a system, one can write balance equations to encompass the experimental observations above. Chemists and physicists are generally interested in the application of the laws of thermodynamics to a change of state in closed systems, while engineers are frequently interested in open systems. For generality, the equations for an open, time-varying system will be written here. The mass balance for the one-component system schematically shown in Fig. 1 is

$$\frac{dM}{dt} = \sum_{j=1}^{N} (\dot{M})_j \tag{1}$$

where M is the total mass of the system at time t, and $(\dot{M})_j$ is the mass flow rate at the jth entry port into the system. For a mixture of C components, the total mass is the sum of the masses of each species i, $M = \sum_{i=1}^{C} M_i$ and $(\dot{M})_j = \sum_{i=1}^{C} (\dot{M}_i)_j$, where $(\dot{M}_i)_j$ is the flow rate of species i at the jth entry point. (Note that the mass balance could also be written on a molar basis; however, since the total number of moles and the number of moles of each species are not conserved on a chemical reaction, that form of the equation is a more complicated.)

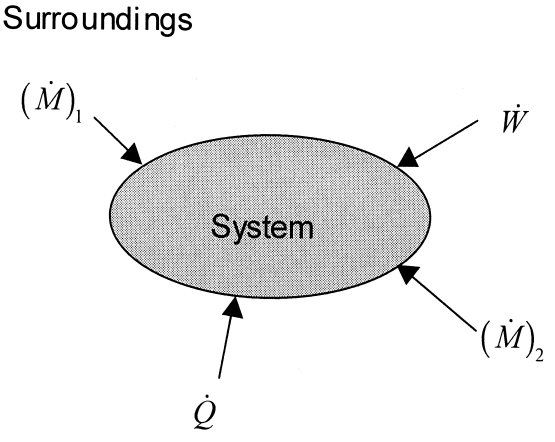


FIGURE 1 A schematic diagram of a system open to the flows of mass, heat, and work.

B. First Law

Using the sign convention that any flow that increases the energy of the system is positive, the energy balance for an open system is

$$\frac{d\left[M\left(\hat{U}+\frac{v^2}{2}+\phi\right)\right]}{dt}=\sum_{j=1}^{N}\left[\dot{M}\left(\hat{H}+\frac{v^2}{2}+\phi\right)\right]_j+\dot{W}+\dot{Q}-P\frac{dV}{dt} \tag{2}$$

The term on the left is the rate of change of the total energy of the system written as a product of the mass of the system and the energy per unit mass. This includes the internal energy $\hat{U}$, the kinetic energy $v^2/2$, and the potential energy ϕ. The first term on the right accounts for the fact that each element of mass entering or leaving the system carries with it its specific enthalpy, $\hat{H}=\hat{U}+P\hat{V}$, the sum of the specific internal energy and energy due to the product of the specific volume and the pressure at the entry port. This term is summed over all entry ports. The remaining terms are the rate at which work is done on the system, $\dot{W}$, by mechanisms that do not involve a change of the system boundaries, referred to as *shaft work*; the rate at which heat or thermal energy enters the system, $\dot{Q}$; and the rate at which work is done on the system by compression or expansion of the system boundaries. A version of this equation that explicitly includes different species in multicomponent mixtures will be considered later. Also, the equation above assumes a constant pressure at the system boundary. If this is not the case, the last term is replaced by an integral over the surface of the system.

C. Second Law

To complete the formulation of thermodynamics, a balance equation is needed for another state property of the system that accounts for such experimental observations as: (1) isolated systems evolve to a state of equilibrium and not in the opposite direction, and (2) while mechanical (kinetic and potential) energy can be completely converted into heat, thermal energy can only partially be converted into mechanical energy, the rest remaining as thermal energy of a lower temperature.

Because mass, energy, and momentum are the only conserved quantities and the momentum balance is of little use in thermodynamics, the additional balance equation will be for a nonconserved property—that is, a property that can be created or destroyed in a change of state.

There are many formulations of the second law of thermodynamics to describe these observations. The one that will be used here states, by postulate, that there is a state function called the *entropy*, denoted by the symbol S (and $\hat{S}$ for entropy per unit mass), with a rate of change given by:

$$\frac{d(M\hat{S})}{dt}=\sum_{j=1}^{N}(\dot{M})_j\hat{S}_j+\frac{\dot{Q}}{T}+\dot{S}_{\text{gen}} \tag{3}$$

where $\dot{S}_{\text{gen}}$, which is greater than or equal to zero, is the rate of entropy generation in a process due to nonuniformities, gradients, and irreversibilities in the system. It is found that $\dot{S}_{\text{gen}}=0$ in a system at equilibrium without any internal flows, and that $\dot{S}_{\text{gen}}$ is greater than zero when such flows occur. The fact that $\dot{S}_{\text{gen}}\geq 0$ and cannot be less than zero encompasses the experimental observations above, as well as many others; indeed, $\dot{S}_{\text{gen}}\geq 0$ is the essence of the second law of thermodynamics. The third law of thermodynamics states that the entropy of all substances in the perfect crystalline state is zero at the absolute zero of temperature. This law is the basis for calculating absolute values of the entropy.

IV. CRITERIA FOR EQUILIBRIUM AND STABILITY

Consider a system that is closed (all $\dot{M}=0$), adiabatic ($\dot{Q}=0$), of constant volume ($dV/dt=0$), without work flows ($\dot{W}$), and stationary (so that there are no changes in kinetic or internal energy). The mass balance, first and second law equations for this system are

$$\frac{dM}{dt}=0;\quad M\frac{d\hat{U}}{dt}=0;\quad M\frac{d\hat{S}}{dt}=\dot{S}_{\text{gen}}\geq 0 \tag{4}$$

The first equation (mass balance) shows that the total mass of this system is constant, and the energy balance (first law) shows that the internal energy per unit mass is constant. The second law (entropy balance) states that the entropy of the system will increase until the system reaches the equilibrium state in which there are no internal flows so that $\dot{S}_{\text{gen}}=0$, and $\frac{d\hat{S}}{dt}=0$; that is, the entropy per unit mass is constant. Now, since $\hat{S}$ is increasing on the approach to equilibrium, and constant equilibrium it follows that the criterion for equilibrium is

$$\hat{S}=\text{maximum}\quad\text{for a system of constant } M, U, \text{ and } V \tag{5a}$$

Mathematically, the equilibrium state is found by observing that for any differential change,

$$d\hat{S}=0\quad\text{for a system of constant } M, U, \text{ and } V$$

and

$$d^2\hat{S}<0 \tag{5b}$$

The first of these equations is used to identify a stationary state of the system, and the second ones ensure that the stationary state is a stable, equilibrium state (that is, a state in which the entropy is a maximum subject to the constraints, and not a minimum).

Similar arguments can be used to identify the mathematical criteria for equilibrium and stability in systems subject to other constraints. Some results are

$$\hat{A} = \hat{U} - T\hat{S} = \text{minimum} \quad (6)$$
$$d\hat{A} = 0 \quad \text{and} \quad d^2\hat{A} > 0$$
$$\text{for a system of constant } M,\ T, \text{ and } V$$

and

$$\hat{G} = \hat{H} - T\hat{S} = \hat{U} + P\hat{V} - T\hat{S} = \text{minimum} \quad (7)$$
$$d\hat{G} = 0 \quad \text{and} \quad d^2\hat{G} > 0$$
$$\text{for a system of constant } M,\ T, \text{ and } P$$

The equations above define the Gibbs free energy G and the Helmholtz free energy A.

From the first of the stability criteria above ($d^2\hat{S} \leq 0$) one can derive that, for a stable equilibrium state to exist for a pure substance, the following criteria must be met:

$$C_V = \left(\frac{\partial \underline{U}}{\partial T}\right)_V > 0 \quad \text{and} \quad \left(\frac{\partial P}{\partial \underline{V}}\right)_T < 0 \quad (8)$$

(In these equations, we have used an underbar to designate a molar property, and C_V is the constant volume heat capacity.) If these criteria are not met, the state is not a stable one, and either another state of aggregation or a two-phase system (i.e., vapor + liquid) is the equilibrium state. The stability criteria for a multicomponent mixture are much more complicated, involving derivatives of the free energy function with respect to composition.

V. PURE COMPONENT PROPERTIES

A. Interrelationships Between State Variables

The first and second laws of thermodynamics are in terms of internal energy and entropy, though the properties that are easiest to measure are temperature and pressure. In order to determine how the properties of a pure substance change with changes in temperature and pressure, consider a stationary, closed system of constant mass without any shaft work. The first and second laws for such a system (on a molar basis) are

$$\frac{d\underline{U}}{dt} = \dot{Q} - P\frac{d\underline{V}}{dt} \quad \text{and} \quad \frac{d\underline{S}}{dt} = \frac{\dot{Q}}{T} + \dot{S}_{\text{gen}} \quad (9)$$

Our interest is in the change of properties between two equilibrium states and, since any convenient path can be used for the calculation, a reversible path is used so that $\dot{S}_{\text{gen}} = 0$. Using this, and combining the two equations above, we obtain:

$$\frac{d\underline{U}}{dt} = T\frac{d\underline{S}}{dt} - P\frac{d\underline{V}}{dt} \quad (10)$$

usually written simply as $d\underline{U} = T\,d\underline{S} - P\,d\underline{V}$. By the chain rule of partial differentiation, one has

$$d\underline{X} = \left(\frac{\partial \underline{X}}{\partial \underline{Y}}\right)_{\underline{Z}} d\underline{Y} + \left(\frac{\partial \underline{X}}{\partial \underline{Z}}\right)_{\underline{Y}} d\underline{Z} \quad (11)$$

From this equation, we find that:

$$\left(\frac{\partial \underline{U}}{\partial \underline{S}}\right)_{\underline{V}} = T; \quad \left(\frac{\partial \underline{U}}{\partial \underline{V}}\right)_{\underline{S}} = -P; \quad \left(\frac{\partial \underline{S}}{\partial \underline{V}}\right)_{\underline{U}} = \frac{P}{T} \quad (12)$$

Two mathematical properties for the partial derivatives of interest here are

$$\left(\frac{\partial \underline{X}}{\partial \underline{Y}}\right)_{\underline{Z}} = \frac{1}{(\partial \underline{Y}/\partial \underline{X})_{\underline{Z}}}$$
$$\left(\frac{\partial \underline{X}}{\partial \underline{Y}}\right)_{\underline{Z}} = \left(\frac{\partial \underline{X}}{\partial \underline{K}}\right)_{\underline{Z}} \left(\frac{\partial \underline{K}}{\partial \underline{Y}}\right)_{\underline{Z}} \quad (13)$$

Using these equations together with Eq. (12) one obtains:

$$\left(\frac{\partial \underline{U}}{\partial \underline{S}}\right)_{\underline{V}} = T = \left(\frac{\partial \underline{U}}{\partial T}\right)_{\underline{V}} \left(\frac{\partial T}{\partial \underline{S}}\right)_{\underline{V}} = C_V \left(\frac{\partial T}{\partial \underline{S}}\right)_{\underline{V}}$$
$$\left(\frac{\partial T}{\partial \underline{S}}\right)_{\underline{V}} = \frac{T}{C_V} \quad \text{or} \quad \left(\frac{\partial \underline{S}}{\partial T}\right)_{\underline{V}} = \frac{C_V}{T} \quad (14)$$

B. Maxwell's Relations

A property of continuous mathematical functions, such as the thermodynamic properties here, is that mixed second derivatives are equal; that is,

$$\left.\frac{\partial}{\partial \underline{X}}\right|_{\underline{Y}} \left(\frac{\partial \underline{Z}}{\partial \underline{Y}}\right)_{\underline{X}} = \left.\frac{\partial}{\partial \underline{Y}}\right|_{\underline{X}} \left(\frac{\partial \underline{Z}}{\partial \underline{X}}\right)_{\underline{Y}} \quad (15)$$

Using this property with Eq. (10), one obtains the following Maxwell relations:

$$\left(\frac{\partial \underline{S}}{\partial \underline{V}}\right)_T = \left(\frac{\partial P}{\partial T}\right)_{\underline{V}}; \quad \left(\frac{\partial \underline{S}}{\partial P}\right)_T = -\left(\frac{\partial \underline{V}}{\partial T}\right)_P;$$
$$\left(\frac{\partial T}{\partial P}\right)_{\underline{S}} = \left(\frac{\partial \underline{V}}{\partial \underline{S}}\right)_P; \quad \left(\frac{\partial T}{\partial \underline{V}}\right)_{\underline{S}} = -\left(\frac{\partial P}{\partial \underline{S}}\right)_{\underline{V}} \quad (16)$$

Now, using the chain rule of partial differentiation and the Maxwell relations, we can write

$$d\underline{S} = \left(\frac{\partial \underline{S}}{\partial T}\right)_{\underline{V}} dT + \left(\frac{\partial \underline{S}}{\partial \underline{V}}\right)_{T} d\underline{V} = \frac{C_V}{T} dT + \left(\frac{\partial P}{\partial T}\right)_{\underline{V}} d\underline{V} \quad (17)$$

In a similar fashion, the following equations are obtained:

$$d\underline{S} = \frac{C_P}{T} dT - \left(\frac{\partial \underline{V}}{\partial T}\right)_{P} dP$$

$$d\underline{U} = C_V dT + \left[T\left(\frac{\partial P}{\partial T}\right)_{\underline{V}} - P\right] d\underline{V} \quad (18)$$

$$d\underline{H} = C_P dT + \left[\underline{V} - \left(\frac{\partial \underline{V}}{\partial T}\right)_{P}\right] dP$$

where $C_P = (\frac{\partial H}{\partial T})_P$ is the constant pressure heat capacity.

C. Equations of State

Two types of information are needed to use the equations above for calculating the changes in thermodynamic properties with a change of state. First is heat capacity data. This information is usually available for each component as a function of temperature for liquids and solids or for the ideal gas state (a gas at such low pressure that interactions between the molecules are of negligible importance). The second type of information needed is an interrelation between pressure, temperature, and specific volume, that is, a volumetric equation of state (EOS). Several examples are given below:

$$P\underline{V} = RT \quad \text{or} \quad Z(T, P) = \frac{P\underline{V}}{RT} = 1 \quad \text{ideal gas EOS}$$

$$P = \frac{RT}{\underline{V} - b} - \frac{a}{\underline{V}^2}$$

or

$$Z(T, P) = \frac{P\underline{V}}{RT} = \frac{\underline{V}}{\underline{V} - b} - \frac{a}{RT\underline{V}} \quad \text{van der Waals EOS}$$

$$P = \frac{RT}{\underline{V} - b} - \frac{a(T)}{\underline{V} \cdot (\underline{V} + b) + b \cdot (\underline{V} - b)} \quad \text{Peng–Robinson EOS}$$

$$Z(T, P) = \frac{P\underline{V}}{RT} = 1 + \frac{B(T)}{\underline{V}} + \frac{C(T)}{\underline{V}^2} + \cdots \quad \text{virial EOS}$$

$$\frac{P\underline{V}}{RT} = Z(T, P) = 1 + \left(B - \frac{A}{RT} - \frac{C}{RT^3}\right)\frac{1}{\underline{V}} + \left(b - \frac{a}{RT}\right)\frac{1}{\underline{V}^2} + \frac{a\alpha}{RT\underline{V}^5} + \frac{\beta}{RT^3\underline{V}}\left(1 + \frac{\gamma}{\underline{V}^2}\right) \times \exp(-\gamma/\underline{V}^2) \quad \text{Benedict–Webb–Rubin EOS}$$

Many other volumetric equations of state have been proposed, including more complicated ones when high accuracy is needed. In these equations, $a(T)$, $B(T)$, and $C(T)$ are functions of temperature; all other parameters are constants specific to each fluid.

The combination of heat capacity data, a volumetric equation of state, and Eqs. (17) and (18) allows the change in thermodynamic properties between any two states to be computed. However, again, a convenient path rather than the actual path is used for the calculation. For example, to compute the change in molar enthalpy between the states $(P_1,\ T_1)$ and $(P_2,\ T_2)$, the path followed is $(P_1,\ T_1) \rightarrow (P = 0,\ T_1) \rightarrow (P = 0,\ T_2) \rightarrow (P_2,\ T_2)$. In this way, the equation of state is used for steps 1 and 3, and the available ideal gas heat capacity is used in step 2:

$$\underline{H}(T_2, P_2) - \underline{H}(T_1, P_1) = \int_{P_1, T_1}^{P=0, T_1} \left[\underline{V} - \left(\frac{\partial \underline{V}}{\partial T}\right)_{P}\right] dP + \int_{P=0, T_1}^{P=0, T_2} C_P dT + \int_{P_0, T_2}^{P_1, T_2} \left[\underline{V} - \left(\frac{\partial \underline{V}}{\partial T}\right)_{P}\right] dP \quad (19)$$

Similar equations are used to compute the change in other thermodynamic properties.

VI. PHASE EQUILIBRIUM IN ONE-COMPONENT SYSTEMS

A. Criterion for Phase Equilibrium

For a one-component open system with no shaft work, the first and second law equations (on a molar basis) are

$$\frac{dU}{dt} = \dot{N}\underline{H} + \dot{Q} - P\frac{dV}{dt} \quad \text{and} \quad \frac{dS}{dt} = \dot{N}\underline{S} + \frac{\dot{Q}}{T} + \dot{S}_{\text{gen}} \quad (20)$$

Again, to compute property changes consider a path on which $\dot{S}_{\text{gen}} = 0$, to obtain:

$$\frac{dU}{dt} = T\frac{dS}{dt} - P\frac{dV}{dt} + \underline{G}\frac{dN}{dt}$$

or simply

$$dU = T\,dS - P\,dV + \underline{G}\,dN \quad (21)$$

Analogous relations are obtained for other thermodynamic properties. For example,

$$dG = \left(\frac{\partial G}{\partial T}\right)_{P,N} dT + \left(\frac{\partial G}{\partial P}\right)_{T,N} dP + \left(\frac{\partial G}{\partial N}\right)_{T,P} dN = -S\,dT + V\,dP + \underline{G}\,dN \quad (22)$$

To obtain the criterion for phase equilibrium in a pure fluid, consider a closed system at constant temperature

and pressure, consisting of two subsystems, I and II, with mass freely transferable between them. As the composite system is closed to external mass flows,

$$dN = dN^{\mathrm{I}} + dN^{\mathrm{II}} = 0 \quad \text{or} \quad dN^{\mathrm{II}} = -dN^{\mathrm{I}} \tag{23}$$

Because the temperature and pressure are fixed and are the same in both subsystems, the change in the Gibbs free energy of the combined system accompanying an interchange of mass is

$$dG = \underline{G}^{\mathrm{I}}\, dN^{\mathrm{I}} + \underline{G}^{\mathrm{II}}\, dN^{\mathrm{II}} \tag{24}$$

At equilibrium, G is a maximum so that $dG = 0$ for all exchanges of mass. Therefore,

$$dG = 0 = \underline{G}^{\mathrm{I}}\, dN^{\mathrm{I}} + \underline{G}^{\mathrm{II}}\, dN^{\mathrm{II}} = (\underline{G}^{\mathrm{I}} - \underline{G}^{\mathrm{II}})\, dN^{\mathrm{I}}$$

This must be true for any value of dN^{I}, so that the condition for phase equilibrium is

$$\underline{G}^{\mathrm{I}}(T, P) = \underline{G}^{\mathrm{II}}(T, P) \quad \text{or equivalently}$$
$$f^{\mathrm{I}}(T, P) = f^{\mathrm{II}}(T, P) \tag{25}$$

where the fugacity, denoted by the symbol f, which is a function of temperature and pressure, is

$$\begin{aligned}\frac{f(T, P)}{P} &= \exp\left[\frac{\underline{G}(T, P) - \underline{G}^{IG}(T, P)}{RT}\right] \\ &= \exp\left[\frac{1}{RT}\int_{P=0}^{P}\left(\underline{V} - \frac{RT}{P}\right) dP\right] \\ &= \exp\left[\frac{1}{RT}\int_{\underline{V}=\infty}^{\underline{V}=ZRT/P}\left[\frac{RT}{\underline{V}} - P\right] d\underline{V}\right. \\ &\quad \left. - \ln Z(T, P) + Z(T, P) - 1\right] \end{aligned} \tag{26}$$

It is easily shown that Eq. (25) is the condition for equilibrium for composite systems subject to other constraints (i.e., closed systems at constant U and V or constant T and V, among others).

B. Calculation of Phase Equilibrium

Figure 2 shows isotherms (lines of constant temperature) on a pressure–volume plot computed using a typical equation of state of the van der Waals form. In this diagram, $T_1 < T_2 < T_3 < T_4 < T_5$. Note that at temperatures T_1 and T_2 there are regions where $(\partial P/\partial \underline{V})_T > 0$, which violates the stability criterion of Eq. (8). Consequently, two phases (a vapor and a liquid) will form in these regions. The thermodynamic properties of the coexisting states are found by requiring that each of the temperature, pressure, and fugacity of both phases be the same. Algorithms and computer codes for such calculations appear in the applied

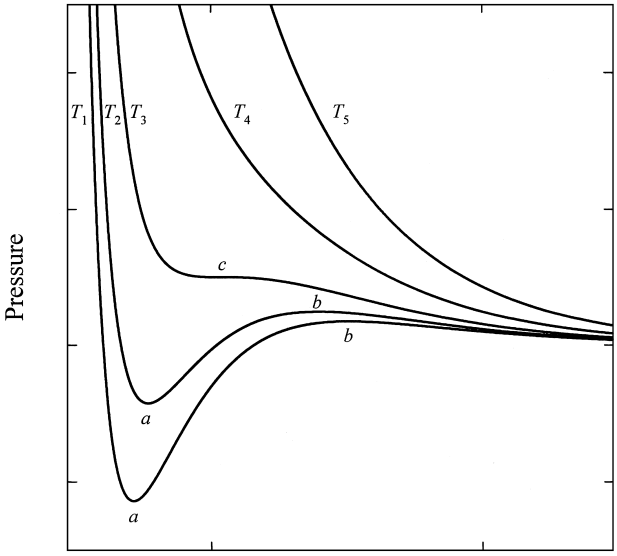


FIGURE 2 *P–V–T* plot for a typical cubic equation of state showing thermodynamically unstable regions (between points *a* and *b*). Point *c* is the critical point.

thermodynamics literature. Figure 3 is a redrawn version of the previous figure replacing the unstable region with the dome-shaped two-phase coexistence region. The left side of the dome gives the liquid properties as a function of the state variables; the vapor properties are given by the

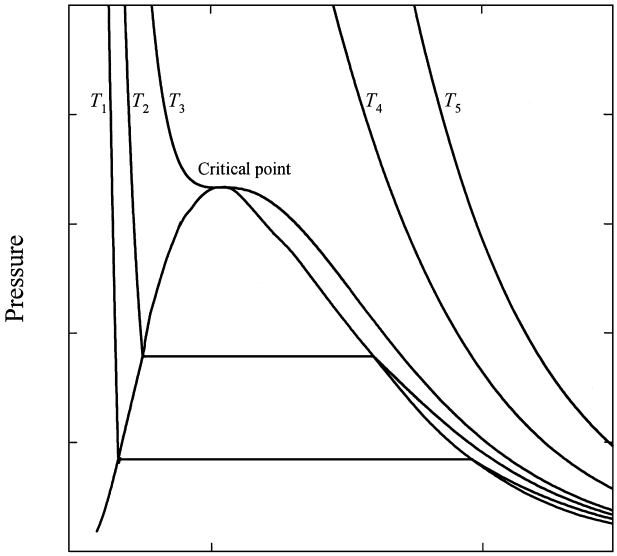


FIGURE 3 *P–V–T* plot for a cubic equation of state with the unstable region replaced with the vapor–liquid equilibrium coexistence region.

right side. A tie line (horizontal line) of constant temperature and pressure connects the two equilibrium phases. The liquid and vapor properties become identical at the peak of the two-phase dome, referred to as the *critical point*, which is point c in Fig. 2. Mathematically, this is the point, at which the equation of state has an inflection point, $(\partial P/\partial \underline{V})_T = (\partial^2 P/\partial \underline{V}^2)_T = 0$, and is a unique point on a pure component phase diagram. The temperature, pressure, and density at the critical point are referred as the *critical temperature*, T_c, the *critical pressure*, P_c, and the *critical volume* V_c, respectively. These conditions are frequently used to determine the values of the parameters in an equation of state.

When an equation of state is not available for a liquid, the fugacity is calculated from:

$$\begin{aligned}\frac{f(T,P)}{P} &= \exp\left[\frac{1}{RT}\int_{P=0}^{P}\left(\underline{V}-\frac{RT}{P}\right)dP\right]\\ &= \exp\left[\frac{1}{RT}\int_{P=0}^{P^{vap}(T)}\left(\underline{V}^{vap}-\frac{RT}{P}\right)dP\right.\\ &\quad\left.+\frac{1}{RT}\int_{P^{vap}(T)}^{P}\left(\underline{V}^{liq}-\frac{RT}{P}\right)dP\right]\\ &= \frac{f(T,P^{vap})}{P^{vap}(T)}\\ &\quad\times\exp\left[\frac{1}{RT}\int_{P^{vap}(T)}^{P}\left(\underline{V}^{liq}-\frac{RT}{P}\right)dP\right]\\ &= \frac{f(T,P^{vap})}{P^{vap}(T)}\frac{P^{vap}(T)}{P}\\ &\quad\times\exp\left[\frac{1}{RT}\int_{P^{vap}(T)}^{P}\underline{V}^{liq}\,dP\right]\end{aligned}$$

or

$$f(T,P) = P^{vap}(T)\frac{f(T,P^{vap})}{P^{vap}(T)} \times\exp\left[\frac{1}{RT}\int_{P^{vap}(T)}^{P}\underline{V}^{liq}\,dP\right] \tag{27}$$

At low vapor and total pressures, this equation reduces to $f(T,P) = P^{vap}(T)$. At higher pressures, the value of the first correction term:

$$\frac{f(T,P^{vap})}{P^{vap}(T)} = \exp\left[\frac{1}{RT}\int_{P=0}^{P^{vap}(T)}\left(\underline{V}^{vap}(T,P)-\frac{RT}{P}\right)dP\right] \tag{28}$$

must be computed; note that this involves the equation of state only for the vapor. Finally, at very high pressures, the exponential term in Eq. (27), known as the *Poynting correction*, is computed using the liquid specific volume.

C. Clapeyron and Clausius–Clapeyron Equations

At equilibrium between phases, the molar Gibbs free energy is the same in both phases, that is $\underline{G}^{\mathrm{I}}(T,P) = \underline{G}^{\mathrm{II}}(T,P)$. For small changes in temperature, the corresponding change in the equilibrium pressure can be computed from:

$$\begin{aligned}d\underline{G}^{\mathrm{I}}(T,P) &= d\underline{G}^{\mathrm{II}}(T,P)\\ \underline{V}^{\mathrm{I}}\,dP - \underline{S}^{\mathrm{I}}dT &= \underline{V}^{\mathrm{II}}dP - \underline{S}^{\mathrm{II}}\,dT\end{aligned}$$

or

$$\left(\frac{dP}{dT}\right)_{\underline{G}^{\mathrm{I}}=\underline{G}^{\mathrm{II}}} = \left(\frac{\underline{S}^{\mathrm{II}}-\underline{S}^{\mathrm{I}}}{\underline{V}^{\mathrm{II}}-\underline{V}^{\mathrm{I}}}\right) = \frac{1}{T}\left(\frac{\underline{H}^{\mathrm{II}}-\underline{H}^{\mathrm{I}}}{\underline{V}^{\mathrm{II}}-\underline{V}^{\mathrm{I}}}\right) = \frac{\Delta\underline{H}}{T\,\Delta\underline{V}} \tag{29}$$

which is the Clapeyron equation. This equation is applicable to vapor–liquid, solid–liquid, solid–vapor, and solid–solid phase transitions. In the case of low-pressure vapor–liquid equilibrium,

$$\Delta\underline{V} = \underline{V}^{vap} - \underline{V}^{liq} \approx \underline{V}^{vap} = \frac{RT}{P}$$

so that

$$\frac{d\ln P^{vap}}{dT} = \frac{\Delta\underline{H}^{vap}}{RT^2}$$

and

$$\ln\frac{P^{vap}(T_2)}{P^{vap}(T_1)} = \int_{T_1}^{T_2}\frac{\Delta\underline{H}^{vap}}{RT^2}\,dT \tag{30}$$

which is the Clausius–Clapeyron equation. For moderate ranges of temperature, where the heat of vaporization can be considered to be approximately constant, this becomes:

$$\ln\frac{P^{vap}(T_2)}{P^{vap}(T_1)} = -\frac{\Delta\underline{H}^{vap}}{R}\left(\frac{1}{T_2}-\frac{1}{T_1}\right) \tag{31a}$$

The simpler form of this equation,

$$\ln P^{vap}(T) = A - \frac{B}{T} \tag{31b}$$

is used as the basis for correlating vapor pressure data.

VII. THERMODYNAMICS OF MIXTURES AND PHASE EQUILIBRIUM

A. Partial Molar Properties

The thermodynamic properties of a mixture are fixed once the values of two state variables (such as temperature and pressure) and the composition of the mixture are fixed. Composition can be specified by either the numbers of moles of all species or the mole fractions of all but one species (as the mole fractions must sum to one). Thus, for example, the change in the Gibbs free energy of a single-phase system of i components is

$$\begin{aligned} dG &= \left(\frac{\partial G}{\partial T}\right)_{P,N} dT + \left(\frac{\partial G}{\partial P}\right)_{T,N} dP \\ &\quad + \sum_{i=1}^{C} \left(\frac{\partial G}{\partial N_i}\right)_{T,P,N_{j\neq i}} dN_i \\ &= -S\,dT + V\,dP + \sum_{i=1}^{C} \bar{G}_i\,dN_i \end{aligned} \tag{32}$$

In this equation, the notation of a partial molar property,

$$\bar{X}_i = \left(\frac{\partial X}{\partial N_i}\right)_{T,P,N_{j\neq i}} = \left(\frac{\partial (N\underline{X})}{\partial N_i}\right)_{T,P,N_{j\neq i}} \tag{33}$$

has been introduced. The partial molar property $\bar{X}_i$ is the amount by which the total system property, X, changes due to the addition of an infinitesimal amount of species i at constant temperature, constant pressure, and constant number of moles of all species except i (designated by $N_{j\neq i}$). A partial molar property is a function not only of species i, but of all species in the mixture and their compositions. Indeed, a major problem in applied thermodynamics is the determination of the partial molar properties.

From Eq. (32) and the first and second laws of thermodynamics, a number of other equations can be derived. Several are listed below:

$$\begin{aligned} dU &= T\,dS - P\,dV + \sum_{i=1}^{C} \bar{G}_i\,dN_i \\ dH &= T\,dS + V\,dP + \sum_{i=1}^{C} \bar{G}_i\,dN_i \\ dA &= -S\,dT - P\,dV + \sum_{i=1}^{C} \bar{G}_i\,dN_i \end{aligned} \tag{34}$$

Note that it is the partial molar Gibbs free energy that appears in each of these equations, which is an indication of its importance in thermodynamics. The partial molar Gibbs free energy of a species, $\bar{G}_i$ is also referred to as the chemical potential μ_i. For simplicity of notation, $\bar{G}_i$ will be used here instead of the more commonly used μ_i.

B. Criteria for Phase and Chemical Equilibrium in Mixtures

Extending the analysis of phase equilibrium used above for a pure fluid to a multicomponent, multiphase system, one obtains as the criterion for equilibrium that,

$$\bar{G}_i^{\mathrm{I}}(T, P, \boldsymbol{x}^{\mathrm{I}}) = \bar{G}_i^{\mathrm{II}}(T, P, \boldsymbol{x}^{\mathrm{II}}) = \bar{G}_i^{\mathrm{III}}(T, P, \boldsymbol{x}^{\mathrm{III}}) = \cdots \tag{35a}$$

or, equivalently,

$$\bar{f}_i^{\mathrm{I}}(T, P, \boldsymbol{x}^{\mathrm{I}}) = \bar{f}_i^{\mathrm{II}}(T, P, \boldsymbol{x}^{\mathrm{II}}) = \bar{f}_i^{\mathrm{III}}(T, P, \boldsymbol{x}^{\mathrm{III}}) = \cdots \tag{35b}$$

where $\boldsymbol{x}$ is being used to indicate the vector of mole fractions of all species present. The fugacity of species i in a mixture $\bar{f}_i$ will be discussed shortly.

Equilibrium in chemical reactions is another important area of chemical thermodynamics. The chemical reaction,

$$\alpha\mathrm{A} + \beta\mathrm{B} + \cdots \Leftrightarrow \rho\mathrm{R} + \sigma\mathrm{S} + \cdots$$

where α, β, etc. are the stoichiometric coefficients will be written as:

$$\rho\mathrm{R} + \sigma\mathrm{S} + \cdots - \alpha\mathrm{A} - \beta\mathrm{B} - \cdots = 0 \tag{36}$$

or simply as

$$\sum_{i=1}^{C} \nu_i \mathrm{I} = 0$$

The mole balance for each species in a chemical reaction can be written using the stoichiometric coefficients in the compact form,

$$N_i = N_{i,0} + \nu_i X \tag{37}$$

where $N_{i,0}$ is the number of moles of species i before any reaction has occurred, and X is the molar extent of reaction, which will have the same value for all species in the reaction. The Gibbs free energy for a closed system at constant temperature and pressure is

$$\begin{aligned} G(T, P, N) &= \sum_{i=1}^{C} N_i \bar{G}_i(T, P, \boldsymbol{N}) \\ &= \sum_{i=1}^{C} (N_i + \nu_i X)\bar{G}_i(T, P, \boldsymbol{N}) \end{aligned} \tag{38}$$

where $\boldsymbol{N}$ is used to indicate the vector of mole numbers of all species present. At equilibrium in a closed system at constant temperature and pressure, G is a maximum, and $dG = 0$. Since the only variation possible is in the molar extent of reaction X, it then follows that for chemical reaction equilibrium,

$$\sum_{i=1}^{C} \nu_i \bar{G}_i(T, P, \boldsymbol{N}) = 0 \quad \text{single chemical reaction,} \tag{39a}$$

In a multiple reaction system, defining ν_{ij} to be the stoichiometric coefficient for species i in the jth reaction, the equilibrium condition becomes:

$$\sum_{i=1}^{C} \nu_{ij} \bar{G}_i(T, P, \boldsymbol{N}) = 0 \quad \text{for each reaction } j = 1, 2, \ldots \tag{39b}$$

In all multiple reaction systems, it is only necessary to consider a set of independent reactions—that is, a reaction set in which no reaction is a linear combination of the others.

Finally, for a system with multiple reactions and multiple phases, the criterion for equilibrium is that Eqs. (35) and (39) must be satisfied simultaneously. That is, for a state of equilibrium to exist in a multiphase, reacting system, each possible process (i.e., transfer of mass between phases or chemical reaction) must be in equilibrium for the system to be in equilibrium. This does not mean that the composition in each phase will be the same.

C. Gibbs Phase Rule

To fix the thermodynamic state of a pure-component, single-phase system, the specification of two state properties is required. Thus, the system is said to have two degrees of freedom, F. To fix the thermodynamic state of a nonreacting, C-component, single-phase system, the values of two state properties and $C-1$ mole fractions are required (the remaining mole fraction is not an independent variable as all the mole fractions must sum to one) for a total of $C+1$ variables. That is, $F = C+1$. Consider a system consisting of C components, P phases, and M independent chemical reactions. Since $C+1$ state properties are needed to fix each phase, it would appear that the system has $P(C+1)$ degrees of freedom. However, since the temperature is the same in all phases, specifying the temperature in one phase fixes its values in the $P-1$ other phases. Similarly, fixing the pressure in one phase sets its values in the $P-1$ remaining phases. That the fugacity of each species must be the same in each phase removes another $C(P-1)$ degrees of freedom. Finally, that the criterion for chemical equilibrium for each of the M independent reactions must be satisfied places another additional M constraints on the system. Therefore, the actual number of degrees of freedom is

$$\begin{aligned} F &= P\cdot(C+1) - (P-1) - (P-1) - C\cdot(P-1) - M \\ &= C - P - M + 2 \end{aligned} \tag{40}$$

This result is the Gibbs phase rule. It is important to note that this gives the number of state properties needed to completely specify the thermodynamic state of each of the phases in the multicomponent, multiphase, multireaction system. However, such a specification does not give information on the relative amounts of the coexisting phases, or the total system size. Such additional information comes from the specification of the initial state and the species mass balances.

VIII. MIXTURE PHASE EQUILIBRIUM CALCULATIONS

Central to the calculation of equilibria in mixtures is the fugacity of species i in the mixture $\bar{f}_i$ which is given by:

$$\begin{aligned} \frac{\bar{f}_i(T, P)}{x_i P} &= \exp\left[\frac{\bar{G}_i(T, P, \boldsymbol{x}) - \bar{G}_i^{IGM}(T, P, \boldsymbol{x})}{RT}\right] \\ &= \exp\left[\frac{1}{RT}\int_{P=0}^{P}\left(\bar{V}_i - \frac{RT}{P}\right)dP\right] \\ &= \exp\left[\frac{1}{RT}\int_{\underline{V}=\infty}^{\underline{V}=ZRT/P}\left[\frac{RT}{\underline{V}} - N\left(\frac{\partial P}{\partial N_i}\right)_{T,V,N_{j\neq i}}\right]d\underline{V} - \ln Z(T, P, \boldsymbol{x})\right] \end{aligned} \tag{41}$$

In this equation, the superscript IGM indicates an ideal gas mixture—that is, a mixture that has the following properties:

$$\begin{aligned} &PV^{IGM} = \left(\sum_{i=1}^{C} N_i\right)RT \text{ or } P\underline{V}^{IGM} = \left(\sum_{i=1}^{C} x_i\right)RT \\ &\quad \text{so that} \quad \bar{V}_i^{IGM}(T, P, \boldsymbol{x}) = \underline{V}_i^{IG}(T, P) = RT/P \\ &\underline{U}^{IGM}(T, P, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{U}_i^{IG}(T, P) \\ &\quad \text{so that} \quad \bar{U}_i^{IGM}(T, P, \boldsymbol{x}) = \underline{U}_i^{IG}(T, P) \\ &\underline{H}^{IGM}(P, T, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{H}_i^{IG}(T, P) \\ &\quad \text{so that} \quad \bar{H}_i^{IGM}(T, P, \boldsymbol{x}) = \underline{H}_i^{IG}(T, P) \\ &\underline{S}^{IGM}(T, P, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{S}_i^{IG}(T, P) - R\sum_{i=1}^{C} x_i \ln x_i \\ &\quad \text{so that} \quad \bar{S}_i^{IGM}(T, P, \boldsymbol{x}) = \underline{S}_i^{IG}(T, P) - R\ln x_i \\ &\underline{A}^{IGM}(T, P, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{A}_i^{IG}(T, P) + RT\sum_{i=1}^{C} x_i \ln x_i \\ &\quad \text{so that} \quad \bar{A}_i^{IGM}(T, P, \boldsymbol{x}) = \underline{A}_i^{IG}(T, P) + RT\ln x_i \end{aligned} \tag{42}$$

$$\underline{G}^{IGM}(T, P, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{G}_i^{IG}(T, P) + RT \sum_{i=1}^{C} x_i \ln x_i$$

$$\text{so that} \quad \bar{G}_i^{IGM}(T, P, \boldsymbol{x}) = \underline{G}_i^{IG}(T, P) + RT \ln x_i$$

Also of interest is the ideal mixture whose properties are given by:

$$V^{IM}(T, P, \boldsymbol{x}) = \sum_{i=1}^{C} N_i \underline{V}_i(T, P)$$

$$\text{so that} \quad \bar{V}_i^{IM}(T, P, \boldsymbol{x}) = \underline{V}_i(T, P)$$

$$\underline{U}^{IM}(T, P, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{U}_i(T, P)$$

$$\text{so that} \quad \bar{U}_i^{IM}(T, P, \boldsymbol{x}) = \underline{U}_i(T, P)$$

$$\underline{H}^{IM}(P, T, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{H}_i(T, P)$$

$$\text{so that} \quad \bar{H}_i^{IM}(T, P, \boldsymbol{x}) = \underline{H}_i(T, P)$$

$$\underline{S}^{IM}(P, T, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{S}_i(T, P) - R \sum_{i=1}^{C} x_i \ln x_i$$

$$\text{so that} \quad \bar{S}_i(T, P, \boldsymbol{x}) = \underline{S}_i(T, P) - R \ln x_i \qquad (43)$$

$$\underline{A}^{IM}(P, T, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{A}_i(T, P) + RT \sum_{i=1}^{C} x_i \ln x_i$$

$$\text{so that} \quad \bar{A}_i^{IM}(T, P, \boldsymbol{x}) = \underline{A}_i(T, P) + RT \ln x_i$$

$$\underline{G}^{IM}(P, T, \boldsymbol{x}) = \sum_{i=1}^{C} x_i \underline{G}_i(T, P) + RT \sum_{i=1}^{C} x_i \ln x_i$$

$$\text{so that} \quad \bar{G}_i^{IM}(T, P, \boldsymbol{x}) = \underline{G}_i(T, P) + RT \ln x_i$$

While the equations for the ideal mixture appear very similar to those for the ideal gas mixture, there are two important distinctions between them. First, the IGM only relates to gaseous mixtures, while the IM is applicable to gases, liquids, and solids. Second, in the IGM the pure component property is that of the ideal gas at the conditions of the mixture, while in the IM the pure component properties are at the same temperature, pressure, and *state of aggregation* of the mixture. Note that in an ideal gas mixture,

$$\bar{V}_i^{IGM}(T, P, \underline{x}) = \frac{RT}{P} \quad \text{so that} \quad \bar{f}_i^{IGM}(T, P, \underline{x}) = x_i P \qquad (44a)$$

while in an ideal mixture,

$$\bar{V}_i^{IM}(T, P, \underline{x}) = \underline{V}_i(T, P)$$

$$\text{so that} \quad \bar{f}_i^{IM}(T, P, \underline{x}) = x_i f_i(T, P) \qquad (44b)$$

That is, in the ideal mixture the fugacity of a component is the product of the mole fraction and the pure component fugacity at the same temperature, pressure, and state of aggregation as the mixture.

A. Equations of State for Mixtures

Few mixtures are ideal gas mixtures, or even ideal mixtures; consequently, there are two ways to proceed. The first method is to use an equation of state; this is the description used for all gaseous mixtures and also for some liquid mixtures, though the latter may be difficult if the chemical functionalities of the species in the mixture are very different. Generally, the same forms of equations of state described earlier are used, though the parameters in the equations are now functions of composition. For the virial equation, this composition dependence is known exactly from statistical mechanics:

$$B(T, \boldsymbol{x}) = \sum_{i=1}^{C} \sum_{j=1}^{C} x_i x_j B_{ij}(T),$$

$$C(T, \boldsymbol{x}) = \sum_{i=1}^{C} \sum_{j=1}^{C} \sum_{k=1}^{C} x_j x_j x_k C_{ijk}(T), \ldots \qquad (45)$$

where the only composition dependence is that shown explicitly. For cubic equations of state, the following mixing rules:

$$a(T, \boldsymbol{x}) = \sum_{i=1}^{C} \sum_{j=1}^{C} x_i x_j a_{ij}(T), \quad b(\boldsymbol{x}) = \sum_{i=1}^{C} \sum_{j=1}^{C} x_i x_j b_{ij} \qquad (46)$$

and combining rules:

$$a_{ij}(T) = \sqrt{a_{ii}(T) a_{jj}(T)}(1 - k_{ij}), \quad b_{ij} = \frac{1}{2}(b_{ii} + b_{jj}) \qquad (47)$$

are used, where the binary interaction parameter k_{ij} is adjusted to give the best fit of experimental data. Other, more complicated mixing rules have been introduced in the last decade to better describe mixtures containing very polar compounds and species of very different functionality. There are additional mixing and combining rules for the multiparameter equations of state, and each is specific to the equation used.

B. Phase Equilibrium Calculations Using an Equation of State

If an equation of state can be used to describe both the vapor and liquid phases of a mixture, it can then be used directly for phase equilibrium calculations based on equating the fugacity of each component in each phase:

$$\bar{f}_i^L(T, P, \boldsymbol{x}) = \bar{f}_i^V(T, P, \boldsymbol{y}) \qquad (48)$$

where the superscripts L and V indicate the vapor and liquid phases, respectively, and $\boldsymbol{x}$ and $\boldsymbol{y}$ are the vectors of their compositions. Algorithms for the computer calculation of this type of phase equilibrium calculation are available elsewhere. Because the vapor and liquid phases of hydrocarbons (together with inorganic gases such as CO_2) are well described by simple equations of state, the oil and gas industry typically does phase equilibrium calculations in this manner. Because of the limited applicability of EOS to the liquid phase of polar mixtures, the method below is commonly used for phase equilibrium calculations in the chemical industry.

C. Excess Properties and Activity Coefficients

A description that can be used for liquid and solid mixtures is based on considering any thermodynamic property to be the sum of the ideal mixture property and a second term, the excess property, that accounts for the mixture being nonideal; that is,

$$\begin{aligned}
\underline{H}(T, P, \boldsymbol{x}) &= \underline{H}^{IM}(T, P, \boldsymbol{x}) + \underline{H}^{ex}(T, P, \boldsymbol{x}) \\
&= \sum_{i=1}^{C} x_i \underline{H}(T, P) + \sum_{i=1}^{C} x_i \bar{H}_i^{ex}(T, P, \boldsymbol{x}) \\
\underline{V}(T, P, \boldsymbol{x}) &= \underline{V}^{IM}(T, P, \boldsymbol{x}) + \underline{V}^{ex}(T, P, \boldsymbol{x}) \\
&= \sum_{i=1}^{C} x_i \underline{V}(T, P) + \sum_{i=1}^{C} x_i \bar{V}_i^{ex}(T, P, \boldsymbol{x}) \\
\underline{G}(T, P, \boldsymbol{x}) &= \underline{G}^{IM}(T, P, \boldsymbol{x}) + \underline{G}^{ex}(T, P, \boldsymbol{x}) \\
&= \sum_{i=1}^{C} x_i \underline{G}(T, P) + RT \sum_{i=1}^{C} x_i \ln x_i \\
&\quad + \sum_{i=1}^{C} x_i \bar{G}_i^{ex}(T, P, \boldsymbol{x})
\end{aligned} \tag{49}$$

where

$$\bar{H}_i^{ex} = \left(\frac{\partial N\underline{H}^{ex}}{\partial N_i}\right)_{T,P,N_{j\neq i}}; \quad \bar{V}_i^{ex} = \left(\frac{\partial N\underline{V}^{ex}}{\partial N_i}\right)_{T,P,N_{j\neq i}}; \quad \bar{G}_i^{ex} = \left(\frac{\partial N\underline{G}^{ex}}{\partial N_i}\right)_{T,P,N_{j\neq i}}; \text{ etc.} \tag{50}$$

Of special interest is the commonly used activity coefficient, γ, which is related to the excess partial molar Gibbs free energy as follows:

$$\bar{G}_i^{ex}(T, P, \boldsymbol{x}) = RT \ln \gamma_i(T, P, \boldsymbol{x})$$

For changes in any mixture property $\theta(T, P, \boldsymbol{N})$ we can write:

$$\begin{aligned}
d\theta(T, P, \boldsymbol{N}) = d(N\underline{\theta}) &= \sum_{i=1}^{C} N_i \, d\bar{\theta}_i + \sum_{i=1}^{C} \bar{\theta}_i \, dN_i \\
&= N\left(\frac{\partial \underline{\theta}}{\partial T}\right)_{P,N} dT + N\left(\frac{\partial \underline{\theta}}{\partial P}\right)_{T,N} dP \\
&\quad + \sum_{i=1}^{C} \bar{\theta}_i \, dN_i
\end{aligned}$$

Subtracting the two forms of the equation, and considering only changes at constant temperature and pressure, this reduces to:

$$\sum_{i=1}^{C} N_i \, d\bar{\theta} = \sum_{i=1}^{C} x_i \, d\bar{\theta}_i = 0 \tag{51a}$$

which for a binary mixture can be written as

$$x_1\left(\frac{\partial \bar{\theta}_1}{\partial x_1}\right)_{T,P} + x_2\left(\frac{\partial \bar{\theta}_2}{\partial x_1}\right)_{T,P} = 0$$

and

$$x_1\left(\frac{\partial \bar{\theta}_1^{ex}}{\partial x_1}\right)_{T,P} + x_2\left(\frac{\partial \bar{\theta}_2^{ex}}{\partial x_1}\right)_{T,P} = 0 \tag{51b}$$

since this equation is satisfied identically for the ideal mixture. Special cases of this equation are

$$x_1\left(\frac{\partial \bar{H}_1^{ex}}{\partial x_1}\right)_{T,P} + x_2\left(\frac{\partial \bar{H}_2^{ex}}{\partial x_1}\right)_{T,P} = 0;$$

$$x_1\left(\frac{\partial \bar{V}_1^{ex}}{\partial x_1}\right)_{T,P} + x_2\left(\frac{\partial \bar{V}_2^{ex}}{\partial x_1}\right)_{T,P} = 0$$

$$x_1\left(\frac{\partial \bar{G}_1^{ex}}{\partial x_1}\right)_{T,P} + x_2\left(\frac{\partial \bar{G}_2^{ex}}{\partial x_1}\right)_{T,P} = x_1\left(\frac{\partial \ln \gamma_1}{\partial x_1}\right)_{T,P} + x_2\left(\frac{\partial \ln \gamma_2}{\partial x_1}\right)_{T,P} = 0 \tag{51c}$$

These equations, forms of the Gibbs–Duhem equation, are useful in obtaining partial molar property information from experimental data and for testing the accuracy of such data. For example, by isothermal heat-of-mixing measurements over a range of concentrations, excess enthalpy data can be obtained as follows. For a binary mixture,

$$\Delta\underline{H}^{mix} = (x_1\bar{H}_1 + x_2\bar{H}_2) - (x_1\underline{H}_1 + x_2\underline{H}_2)$$

and

$$\left(\frac{\partial \Delta \underline{H}^{mix}}{\partial x_1}\right)_{T,P} = (\bar{H}_1 - \underline{H}_1) + x_1\left(\frac{\partial \bar{H}_1}{\partial x_1}\right)_{T,P} - (\bar{H}_2 - \underline{H}_2) + x_1\left(\frac{\partial \bar{H}_2}{\partial x_1}\right)_{T,P}$$

Using the Gibbs–Duhem equation and combining the two equations above give:

$$\Delta \underline{H}^{mix} - x_1\left(\frac{\partial \Delta \underline{H}^{mix}}{\partial x_1}\right)_{T,P} = \bar{H}_2 - \underline{H}_2$$

and

$$\Delta \underline{H}^{mix} + x_2\left(\frac{\partial \Delta \underline{H}^{mix}}{\partial x_1}\right)_{T,P} = \bar{H}_1 - \underline{H}_1 \quad (53)$$

Consequently, by having $\Delta \underline{H}^{mix}$ data as a function of composition so that the compositional derivatives can be evaluated, the partial molar enthalpies of each of the species at each composition can be obtained. If the $\Delta \underline{H}^{mix}$ data have been fitted to an equation, usually a polynomial in mole fraction, this can be done analytically. The graphical procedure shown in Fig. 4 can also be used, where the intercepts A and B then give the difference between the partial molar and pure component enthalpies at the indicated concentration. Similar procedures can be used to obtain partial molar volume data from volume change on mixing data. From vapor–liquid equilibrium data, as will be described later, activity coefficient (excess Gibbs free energy) data can be obtained. Also, if partial molar property data have been obtained experimentally, they can be tested for thermodynamic consistency by using the Gibbs–Duhem equation either differentially on a point-by-point basis or by integration over the whole dataset.

Algebraic expressions are generally used to fit excess property data as a function of composition. For example, when the two-parameter expression,

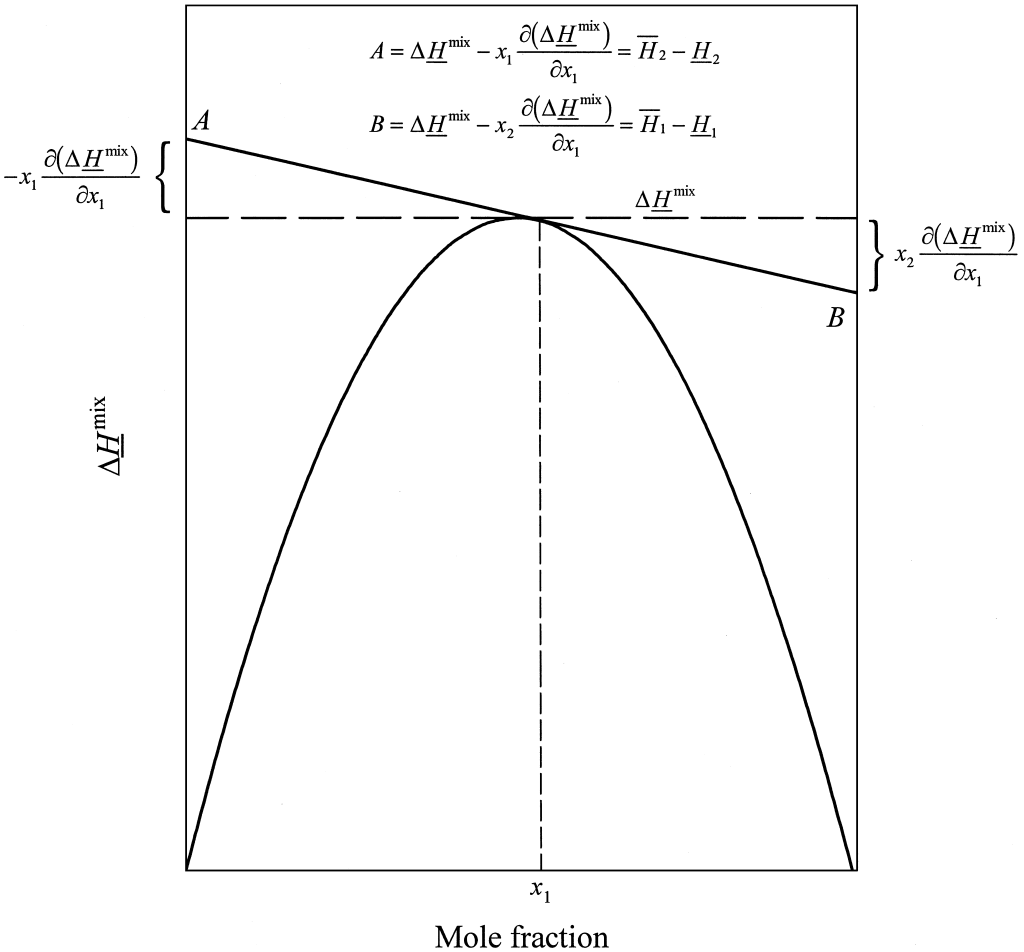


FIGURE 4 Construction illustrating how the difference between the partial molar and pure-component enthalpies can be obtained graphically at a fixed composition from a plot of $\Delta \underline{H}^{mix}$ versus composition in a binary mixture.

$$\underline{\theta}^{ex}(T, P, \boldsymbol{x}) = \frac{ax_1x_2}{(x_1 + x_2b)} \tag{54a}$$

is used, one obtains, in general,

$$\bar{\theta}_1^{ex} = \frac{abx_2^2}{(x_1 + x_2b)^2} \quad \text{and} \quad \bar{\theta}_2^{ex} = \frac{ax_1^2}{(x_1 + x_2b)^2} \tag{54b}$$

and, in particular,

$$\bar{G}_1^{ex} = RT \ln \gamma_1 = \frac{abx_2^2}{(x_1 + x_2b)^2}$$

and

$$\bar{G}_2^{ex} = RT \ln \gamma_2 = \frac{ax_1^2}{(x_1 + x_2b)^2} \tag{54c}$$

which is the Van Laar model. There are many other, and more accurate, activity coefficient models in the thermodynamic literature that are used by chemists and engineers.

D. Phase Equilibrium Calculations Using Activity Coefficients

With this definition of the partial molar excess Gibbs free energy and the activity coefficient, the fugacity of a species in a liquid mixture can be computed from:

$$\bar{f}_i^L(T, P, \boldsymbol{x}) = x_i\gamma_i(T, P, \boldsymbol{x}) f_i^L(T, P) \tag{55}$$

where the fugacity of the pure component is equal to the vapor pressure of the pure component, $P^{vap}(T)$, if the vapor pressure and total pressure are low. If the vapor pressure is above ambient, then the fugacity at this pressure contains a correction that can be computed from the equation of state for the vapor. Also, if the total pressure is much above the pure component vapor pressure, a Poynting correction is added:

$$f_i^L(T, P) = P_i^{vap}(T)\left(\frac{f_i^L\left(T, P_i^{vap}\right)}{P_i^{vap}}\right) \times \exp\left(\int_{P_i^{vap}(T)}^{P} \frac{\underline{V}^L}{RT}\, dP\right) \tag{56}$$

The calculation of vapor–liquid equilibrium using activity coefficient models is then based on:

$$\begin{aligned}\bar{f}_i^L(T, P, \boldsymbol{x}) &= x_i\gamma_i(T, P, \boldsymbol{x}) f_i^L(T, P) \\ &= x_i\gamma_i(T, P, \boldsymbol{x}) P_i^{vap}(T)\left(\frac{f_i^L\left(T, P_i^{vap}\right)}{P_i^{vap}}\right) \\ &\quad \times \exp\left(\int_{P_i^{vap}(T)}^{P} \frac{\underline{V}^L}{RT}\, dP\right) = \bar{f}_i^V(T, P, \boldsymbol{y})\end{aligned} \tag{57}$$

A common application of this equation is to vapor–liquid equilibrium at low pressures, where the vapor can be considered to be an ideal gas mixture and all pressure corrections can be neglected. This leads to the simple equation,

$$x_i\gamma_i(T, \boldsymbol{x}) P_i^{vap}(T) = y_i P \tag{58}$$

relating the compositions of the vapor and liquid phases. If vapor–liquid phase equilibrium data are available, this equation can be used to obtain values of $\gamma_i(T, \boldsymbol{x})$ and, therefore, $\bar{G}_i^{ex}(T, \boldsymbol{x})$ and $\underline{G}^{ex}(T, \boldsymbol{x}) = \sum x_i \bar{G}_i^{ex}$ $(T, \boldsymbol{x}) = RT \sum x_i \ln \gamma_i(T, \boldsymbol{x})$. Alternatively, if activity coefficient or $\underline{G}^{ex}$ data are available or can be predicted, the compositions of the equilibrium phases can be computed. Note that for the case of an ideal solution ($\gamma_i = 1$ for all compositions), the low-pressure vapor–liquid equilibrium relation becomes:

$$x_i P_i^{vap}(T) = y_i P \tag{59a}$$

Also, summing over all species, one then obtains for the ideal solution at low pressure:

$$P(T, \underline{x}) = \sum_{i=1}^{C} x_i P_i^{vap}(T)$$

and

$$y_i = \frac{x_i P_i^{vap}(T)}{P(T, \underline{x})} = \frac{x_i P_i^{vap}(T)}{\sum_{j=1}^{C} x_j P_j^{vap}(T)} \tag{59b}$$

(since $\sum_{i=1}^{c} y_i = 1$). The first of these equations indicates that the total pressure is a linear function of liquid-phase mole fraction. This is known as Raoult's law. The second equation establishes that the vapor and liquid compositions in an ideal solution will be different (except if, fortuitously, the vapor pressures of the components are equal).

The comparable equations for a nonideal mixture at low pressure are

$$P = \sum_{i=1}^{C} x_i\gamma_i(T, \underline{x}) P_i^{vap}(T)$$

and

$$y_i = \frac{x_i\gamma_i(T, \underline{x}) P_i^{vap}(T)}{\sum_{j=1}^{C} x_j\gamma_j(T, \underline{x}) P_j^{vap}(T)} \tag{60a}$$

Figure 5 shows the pressure versus mole fraction behavior for various mixtures. In this figure, curve 1 is for an ideal solution (i.e., Raoult's law). Curves 2 and 3 correspond to solutions with positive deviations from Raoult's law as a result of the activity coefficients of both species being greater than unity. Curves 4 and 5 are similar for the case of negative deviations from Raoult's law ($\gamma < 1$). Figure 6 is a plot of the vapor-phase mole fraction, y, versus the liquid phase mole fraction, x, for these cases. The dashed line in the figure is $x = y$.

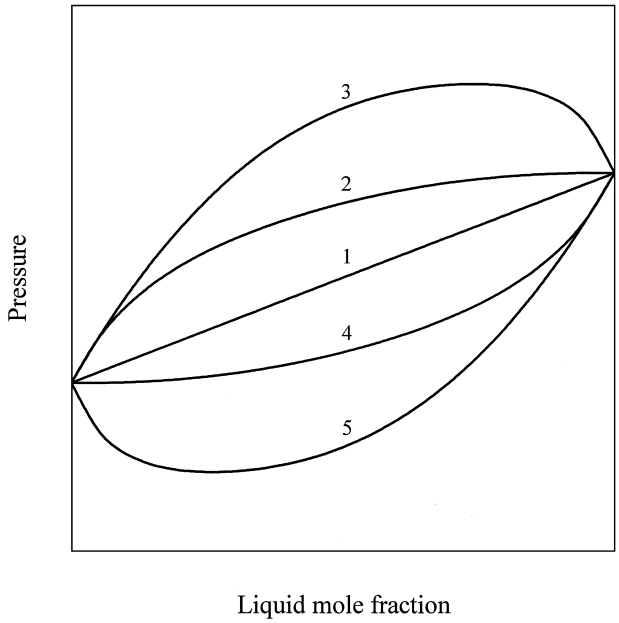


FIGURE 5 Pressure versus liquid composition curves for vapor–liquid equilibrium in a binary mixture. Curve 1 is for an ideal mixture (Raoult's Law). Curves 2 and 3 are for nonideal solutions in which the activity coefficients are greater than unity, and curves 4 and 5 are for nonideal solutions in which the activity coefficients are less than unity. Curves 3 and 5 are for mixtures in which the solution nonideality is sufficiently great as to result in an azeotrope.

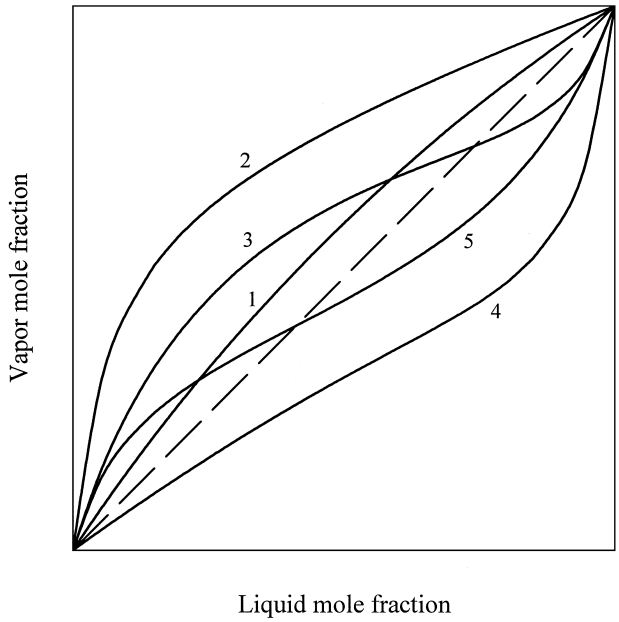


FIGURE 6 Liquid composition versus vapor composition (x vs. y) curves for the mixtures in Fig. 5. The dashed line is the line of $x = y$, and the point of crossing of this line is the azeotropic point.

Curve 3 in Fig. 5 is a case in which the nonideality is sufficiently great that there is a maximum in the pressure versus liquid composition curve. Mathematically, it can be shown that at this maximum the vapor and liquid compositions are identical. This is seen as a crossing of the $x = y$ line in Fig. 6. Such a point is referred to as an *azeotrope*. Curve 5 is another example of a mixture having an azeotrope, although as a result of large negative deviations from Raoult's law. Azeotropes occur as a result of solution nonidealities and are most likely to occur in mixtures of chemically dissimilar species with vapor pressures that are reasonably close. An azeotrope in a binary mixture occurs if:

$$\gamma_1(T, x_1) = \frac{P}{P_1^{vap}(T)} = \frac{x_1 P_1^{vap}(T) + x_2 P_2^{vap}(T)}{P_1^{vap}(T)}$$

and

$$\gamma_2(T, x_1) = \frac{P}{P_2^{vap}(T)} \tag{61}$$

If the azeotropic point of a mixture and the pure component vapor pressures have been measured, the two concentration-dependent activity coefficients can be calculated at this composition. This information can then be used to obtain values of the parameters in a two-parameter activity coefficient model, such as the Van Laar model discussed earlier, and then to predict values of the activity coefficients and the vapor–liquid equilibria over the whole concentration range. The occurrence of azeotropes in multicomponent mixtures is not very common. Calculations for nonideal mixtures at high pressures are considerably more complicated and are discussed in books on applied thermodynamics.

E. Henry's Law

There is an important complication that arises in the calculation of phase equilibrium with activity coefficients: To use Eq. (55) one must be able to calculate the fugacity of the pure component as a liquid at the temperature and pressure of the mixture. This is not possible, for example, if the dissolved component exists only as a gas (i.e., O_2, CO_2, etc.) or as a solid (i.e., sugar, a long-chain hydrocarbon, etc.) as a pure component at the mixture conditions. If the temperature and pressure are not very far from the melting point of the solid or boiling point of the gaseous species, Eq. (27) can still be used by extrapolation of the liquid fugacity (or vapor pressure) into the solid or gaseous states as appropriate. (Such a problem does not arise when using an equation of state, as the species fugacity in a mixture is calculated directly, not with respect to a pure component state.)

If extrapolation over a very large temperature range would be required, a different procedure is used. In this case, Eq. (53) is be replaced by:

$$\bar{f}_i^L(T, P, \underline{x}) = x_i \gamma_i^*(T, P, \underline{x}) \mathcal{H}_i(T, P) \quad (62a)$$

or

$$\bar{f}_i^L(T, P, \underline{x}) = M_i \gamma_i^{\oplus}(T, P, M) \mathcal{H}_i(T, P) \quad (62b)$$

depending on the concentration units used. In these two equations, forms of Henry's law, the fugacity of a gaseous or solid component dissolved in a liquid is calculated based on extrapolation of its behavior when it is highly diluted. In the first equation, the initially linear dependence of the species fugacity at high dilution is used to find the Henry's law constant $\mathcal{H}_i$. Then, the nonlinear behavior at higher concentrations is accounted for by the composition-dependent activity coefficient γ_i^*. In this description, the Henry's law constant depends on temperature and the solvent–solute pair. Also, normalization of the activity coefficient γ_i^* is different from the activity coefficient used heretofore in that its value is unity when the species is infinitely dilute, while $\gamma_i = 1$ in the pure component limit. The relation between the two is

$$\gamma_i^*(x_i) = \frac{\gamma_i(x_i)}{\gamma_i(x_i = 0)} \quad (63)$$

The second form of Henry's law, Eq. (62b), is similar but based on using molality as the concentration variable.

Both types of Henry's law coefficients are generally determined from experiment. Once values are known as a function of temperature, solvent, and solute, the phase behavior involving a solute described by Henry's law can be calculated. For example, at low total pressure, we have for the vapor–liquid equilibrium of such a component:

$$x_i \gamma_i^*(T, P, \underline{x}) \mathcal{H}_i(T, P) = y_i P = P_i$$

or

$$M_i \gamma_i^{\oplus}(T, P, M) \mathcal{H}_i(T, P) = y_i P = P_i \quad (64)$$

depending on the concentration variable used. At higher pressures, a Poynting correction would have to be added to the left side of both equations, and the partial pressure of the species in the vapor phase, P_i, would be replaced by its fugacity, normally calculated from an equation of state.

IX. CHEMICAL EQUILIBRIUM

The calculation of chemical equilibrium is based on Eq. (39). While the partial molar Gibbs free energy or chemical potential of each species in the mixture is needed for the calculation, what is typically available is the Gibbs free energy of formation $\Delta \underline{G}_f$ and the heat (enthalpy) of formation $\Delta \underline{H}_f$ of the pure components from their elements, generally at 25°C and 1 bar. To proceed, one writes:

$$\begin{aligned}\bar{G}_i(T, P, \boldsymbol{x}) &= \underline{G}_i(T, P = 1\,\text{bar}) \\ &\quad + [\,\bar{G}_i(T, P, \boldsymbol{x}) - \underline{G}_i(T, P = 1\,\text{bar})] \\ &= \underline{G}_i(T, P = 1\,\text{bar}) + RT \ln \frac{\bar{f}_i(T, P, \boldsymbol{x})}{f_i(T, P = 1\,\text{bar})} \end{aligned} \quad (65)$$

Then, Eq. (39) can be written as:

$$\begin{aligned}\sum_{i=1}^{C} \nu_i \bar{G}_i(T, P, \boldsymbol{x}) &= \sum_{i=1}^{C} \nu_i \Bigg[\underline{G}_i(T, P = 1\,\text{bar}) \\ &\quad + RT \ln \frac{\bar{f}_i(T, P, \boldsymbol{x})}{f_i(T, P = 1\,\text{bar})} \Bigg] = 0 \end{aligned} \quad (66)$$

Common notation is to define the activity of each species as:

$$a_i(T, P, \boldsymbol{x}) \frac{\bar{f}_i(T, P, \boldsymbol{x})}{f_i(T, P = 1\,\text{bar})} \quad (67)$$

and to define a chemical equilibrium constant $K(T)$ from:

$$\begin{aligned} RT \ln K(T) &= -\sum_{i=1}^{C} \nu_i \underline{G}_i(T, P = 1\,\text{bar}) \\ &= -\sum_{i=1}^{C} \nu_i \Delta \underline{G}_{f,i}(T, P = 1\,\text{bar}) = -\Delta G^{\circ}_{rxn}(T) \end{aligned} \quad (68)$$

leading to:

$$K(T) = \prod_{i=1}^{C} \left(\frac{\bar{f}_i(T, P, \boldsymbol{x})}{f_i(T, P = 1\,\text{bar})} \right)^{\nu_i} = \prod_{i=1}^{C} [a_i(T, P, \boldsymbol{x})]^{\nu_i} \quad (69)$$

where $\Delta G^{\circ}_{rxn}(T)$ is the standard free energy of reaction—that is, the Gibbs free energy change that would occur between reactants in the pure component state to produce products, also as pure components. At 25°C,

$$\begin{aligned} RT \ln K(T &= 25^{\circ}\text{C}) \\ &= -\sum_{i=1}^{C} \nu_i \underline{G}_i(T = 25^{\circ}\text{C}, P = 1\,\text{bar}) \\ &= -\sum_{i=1}^{C} \nu_i \Delta \underline{G}_{f,i}(T = 25^{\circ}\text{C}, P = 1\,\text{bar}) \\ &= -\Delta G^{\circ}_{rxn}(T = 25^{\circ}\text{C}) \end{aligned} \quad (70)$$

Also, the standard heat of reaction is

$$\begin{aligned} \Delta H^{\circ}_{rxn}(T &= 25^{\circ}\text{C}) \\ &= \sum_{i=1}^{C} \nu_i \underline{H}_i(T = 25^{\circ}\text{C}, P = 1\,\text{bar}) \\ &= \sum_{i=1}^{C} \nu_i \Delta \underline{H}_{f,i}(T = 25^{\circ}\text{C}, P = 1\,\text{bar}) \end{aligned}$$

and

$$\Delta H^{\circ}_{rxn}(T) = \Delta H^{\circ}_{rxn}(T = 25^{\circ}\mathrm{C}) + \int_{T=25^{\circ}\mathrm{C}}^{T} \sum_{i=1}^{C} \nu_i \mathrm{C}_{\mathrm{P},i}(T)\, dT \quad (71)$$

Then, using:

$$\frac{\partial}{\partial T}\left(\frac{G}{T}\right)_P = -\frac{H}{T^2} \quad \text{leads to}$$

$$\left(\frac{\partial \ln K(T)}{\partial T}\right)_P = \frac{\Delta H^{\circ}_{rxn}(T)}{T^2} \quad (72a)$$

and

$$\ln \frac{K(T)}{K(T = 25^{\circ}\mathrm{C})} = \int_{T=25^{\circ}\mathrm{C}}^{T} \frac{\Delta H^{\circ}_{rxn}(T)}{RT^2}\, dT = \frac{\Delta H^{\circ}_{rxn}(T = 25^{\circ}\mathrm{C})}{R}\left(\frac{1}{T} - \frac{1}{298.15}\right) + \int_{T=25^{\circ}\mathrm{C}}^{T} \frac{\left[\int_{T^1=25^{\circ}\mathrm{C}}^{T} \sum_{i=1}^{C} \nu_i C_{\mathrm{P},i}(T^1)\, dT^1\right]}{RT^2}\, dT \quad (72b)$$

For a liquid species at low and moderate pressure, and with the pure-component standard state, the activity is

$$a_i(T, P, \boldsymbol{x}) = \frac{\bar{f}_i^L(T, P, \boldsymbol{x})}{f_i^L(T, P = 1\,\mathrm{bar})} = \frac{x_i \gamma_i(T, P, \boldsymbol{x}) f_i^L(T, P)}{f_i^L(T, P = 1\,\mathrm{bar})} = x_i \gamma_i(T, P, \boldsymbol{x}) \quad (73a)$$

The activity of species in the vapor is

$$a_i(T, P, \boldsymbol{y}) = \frac{\bar{f}_i^V(T, P, \boldsymbol{y})}{f_i^V(T, P = 1\,\mathrm{bar})} = \frac{y_i P}{1\,\mathrm{bar}} \quad (73b)$$

where the term on the right of the expression is correct only for an ideal gas mixture. Thus, for example, the chemical equilibrium relation for the low-pressure gas-phase reaction, $\mathrm{H_2} + \frac{1}{2}\mathrm{O_2} \leftrightarrow \mathrm{H_2O}$ is

$$K(T) = \frac{a_{\mathrm{H_2O}}}{a_{\mathrm{H_2}} a_{\mathrm{O_2}}^{1/2}} = \frac{\dfrac{y_{\mathrm{H_2O}} P}{1\,\mathrm{bar}}}{\dfrac{y_{\mathrm{H_2}} P}{1\,\mathrm{bar}}\left(\dfrac{y_{\mathrm{O_2}} P}{1\,\mathrm{bar}}\right)^{1/2}} = \frac{y_{\mathrm{H_2O}}}{y_{\mathrm{H_2}}\left(y_{\mathrm{O_2}}\right)^{1/2}}\left(\frac{1\,\mathrm{bar}}{P}\right)^{1/2} \quad (74)$$

which indicates that as the pressure increases, the conversion of hydrogen and oxygen to water is favored. The equilibrium relation for the low-pressure hydrogenation of benzene to cyclohexane involving hydrogen gas and liquid benzene and cyclohexane $\mathrm{C_6H_6} + 3\mathrm{H_2} \leftrightarrow \mathrm{C_6H_{12}}$ is

$$K(T) = \frac{a_{\mathrm{C_6H_{12}}}}{a_{\mathrm{C_6H_6}} a_{\mathrm{H_2}}^3} = \frac{\dfrac{x_{\mathrm{C_6H_{12}}} \gamma_{\mathrm{C_6H_{12}}} f^L_{\mathrm{C_6H_{12}}}}{f^L_{\mathrm{C_6H_{12}}}}}{\dfrac{x_{\mathrm{C_6H_6}} \gamma_{\mathrm{C_6H_{12}}} f^L_{\mathrm{C_6H_6}}}{f^L_{\mathrm{C_6H_6}}}\left(\dfrac{y_{\mathrm{H_2}} P}{1\,\mathrm{bar}}\right)^3} = \frac{x_{\mathrm{C_6H_{12}}} \gamma_{\mathrm{C_6H_{12}}}}{x_{\mathrm{C_6H_6}} \gamma_{\mathrm{C_6H_6}}}\left(\frac{1\,\mathrm{bar}}{y_{\mathrm{H_2}} P}\right)^3 = \frac{x_{\mathrm{C_6H_{12}}}}{x_{\mathrm{C_6H_6}}}\left(\frac{1\,\mathrm{bar}}{P_{\mathrm{H_2}}}\right)^3 \quad (75)$$

where in the last term in this equation the activity coefficients have been omitted, as benzene and cyclohexane are so chemically similar that they are expected to form an ideal solution, and $P_{\mathrm{H_2}} = y_{\mathrm{H_2}} P$ is the partial pressure of hydrogen in the gas phase.

If the reaction system is closed, then the equilibrium relations have to be solved together with the mass balances. For example, suppose three moles of hydrogen and one mole of oxygen are being reacted to form water. The mass balances for this reaction give:

Species	Initial moles	Moles at equilibrium	Equilibrium mole fraction
$\mathrm{H_2}$	3	$3 - X$	$\dfrac{3 - X}{4 - 0.5X}$
$\mathrm{O_2}$	1	$1 - 0.5X$	$\dfrac{1 - 0.5X}{4 - 0.5X}$
$\mathrm{H_2O}$	0	X	$\dfrac{X}{4 - 0.5X}$
Total moles		$4 - 0.5X$	

The chemical equilibrium relation to be solved for the molar extent of reaction X is, then,

$$K(T) = \frac{y_{\mathrm{H_2O}}}{y_{\mathrm{H_2}}\left(y_{\mathrm{O_2}}\right)^{1/2}}\left(\frac{1\,\mathrm{bar}}{P}\right)^{1/2} = \frac{\dfrac{X}{4 - 0.5X}}{\dfrac{3 - X}{4 - 0.5X}\left(\dfrac{1 - 0.5X}{4 - 0.5X}\right)^{1/2}}\left(\frac{1\,\mathrm{bar}}{P}\right)^{1/2} = \frac{X(4 - 0.5X)^{1/2}}{(3 - X)(1 - 0.5X)^{1/2}}\left(\frac{1\,\mathrm{bar}}{P}\right)^{1/2}$$

Therefore, once the temperature is specified so that value of $K(T)$ can be computed, and the pressure is fixed, the equilibrium molar extent of reaction X can be computed, and from that each of the equilibrium mole fractions.

When several reactions occur simultaneously, a similar procedure is followed in that a chemical equilibrium relation is written for each of the independent reactions, and mass balances are used for each component. The solution

can be complicated since all the reactions are coupled through the mass balances; that is, the molar extent for each reaction will appear in some or all of the equilibrium relations.

When there are many reactions possible, or when there is combined chemical and phase equilibrium, calculation by direct Gibbs free energy minimization may be a better way to proceed. In this method, expressions are written for the partial molar Gibbs free energy of every component in every possible phase (which will involve the mole fractions of all species in that phase), and then a search method is used to find the state of minimum Gibbs free energy (if temperature and pressure are fixed) subject to the mass balance constraints. That is, one identifies the state in which the total Gibbs free energy is a minimum directly, rather than using chemical equilibrium constants.

X. ELECTROLYTE SOLUTIONS

Electrolyte solutions are fundamentally different from the other mixtures so far considered. One reason is that the species, such as salts, ionize in solution so that the nature of the pure component and the substance in solution is very different. Another reason is that, because the ions are charged, the interactions are much stronger and longer range than among molecules. Consequently, the solutions are much more nonideal, and the activity coefficient models used for molecules, such as the simple Van Laar model, are not applicable. Also, the anions and cations originating from a single ionizable substance are present in a fixed ratio.

Consider the ionization reaction $A_{\nu_+}B_{\nu_-} = \nu_+ A^{z_+} + \nu_- B^{z_-}$. Since the initial molecule has no net charge, we have

$$\nu_+ z_+ + \nu_- z_- = 0 \tag{76a}$$

or, on a molar basis,

$$\nu_+ N_A + \nu_- N_B = 0 \tag{76b}$$

where ν_+ and ν_- are the stoichiometric coefficients of the ions A and B in the molecule, and z_+ and z_- are their charges. By Eq. (76b) the number of moles of each ion cannot be changed independently, so the partial molar Gibbs free energy of each ion cannot be separately measured. As the total molar concentration of salt can be varied, the customary procedure is to define a mean ionic activity coefficient $\gamma_\pm$ based on Henry's law, applicable to both ions, and referenced to a hypothetical ideal one-molal solution as follows:

$$\bar{G}_{AB}(T, P, M) = \bar{G}_{AB}^{Ideal}(T, P, M = 1) + \nu RT \ln\left[\frac{M_\pm \gamma_\pm}{M = 1}\right] \tag{77}$$

where $\nu = \nu_+ + \nu_-$ and $M_\pm^\nu = M_A^{\nu_+} M_B^{\nu_-}$ is the mean ionic molality. At very low ionic concentrations, the mean ionic activity coefficient $\gamma_\pm$ can be computed from the Debye–Hückel limiting law:

$$\ln \gamma_\pm = -\alpha |z_+ z_-| \sqrt{I} \tag{78}$$

where

$$I = \frac{1}{2} \sum_{i=\text{ions}} z_i^2 M_i$$

In this equation, I is the ionic strength, the sum is over all ions in solution, and α is a temperature-dependent parameter whose value is 1.178 $(\text{mol/L})^{-0.5}$ for water at 25°C. At higher ionic strengths, the following empirical extensions to the limiting law have been used:

$$\ln \gamma_\pm = -\frac{\alpha |z_+ z_-| \sqrt{I}}{1 + \beta \sqrt{I}}$$

and

$$\ln \gamma_\pm = -\frac{\alpha |z_+ z_-| \sqrt{I}}{1 + \beta \sqrt{I}} + \delta I \tag{79}$$

where $\beta = 1.316$ $(\text{mol/L})^{-0.5}$ for water at 25°C, and δ is an adjustable parameter fit to experimental data. Note that Eq. (78) and the first of Eq. (79) predict a steep and continuing decrease of $\gamma_\pm$ with increasing ionic strength, while the last of Eq. (79) correctly predicts first a decrease in $\gamma_\pm$ and then an increase with increasing ionic strength.

Since a solvent of high dielectric constant is needed for a salt to ionize, ions are not found in the vapor phase at normal conditions. However, the strong nonideality of an electrolyte solution containing ions affects vapor–liquid and reaction equilibria. For example, silver chloride is only very slightly soluble in water. The equilibrium constant for the reaction $AgCl \rightarrow Ag^+ + Cl^-$ is

$$K = \frac{a_{Ag^+} a_{Cl^-}}{a_{AgCl}} = \frac{\dfrac{M_{Ag^+}}{M=1}\dfrac{M_{Cl^-}}{M=1}(\gamma_\pm)^2}{1} = M_{Ag^+} M_{Cl^-} (\gamma_\pm)^2$$

so that

$$M_{Ag^+} = \frac{K}{(\gamma_\pm)^2 M_{Cl^-}} \tag{80}$$

The molality of the silver ion that will dissolve is affected by the addition of other ions. If a salt containing neither silver or chloride ions (e.g., KNO_3) is added to a silver chloride solution, the ionic strength of the solution will increase; this will result in a decrease in the mean ionic activity coefficient at low total ionic strength and an increase in the solubility of Ag^+ ions. Conversely, at higher ionic strength, the mean ionic activity coefficient

will increase, producing a decrease in the solubility of Ag^+ ions. However, if a salt containing a Cl^- ion is added, there will be a small ionic strength effect, but a large common-ion effect resulting in a decrease in the concentration of Ag^+ ions and the solubility of AgCl. That is, because the value of the equilibrium constant is fixed, increasing the Cl^- ion concentration by addition of a Cl-containing salt will depress the Ag^+ ion concentration.

XI. COUPLED REACTIONS

For a state of equilibrium at constant temperature and pressure, the Gibbs free energy should be a minimum. If several chemical reactions occur in a system that are only linked through mass balances, then those reactions that reduce the Gibbs free energy of the system will occur, and those that increase G will not occur. There are, however, other reactions that are more closely coupled. One example is an electrolytic battery in which two electrochemical reactions occur, one of which increases the Gibbs free energy of the system while the other decreases it. When the two half cells are connected, if the sum of the two Gibbs free energy changes is negative, both reactions will occur, including the half-cell reaction that increases the Gibbs free energy system. That is, the reaction with a negative Gibbs free energy change is driving the one with a positive change.

Another example is the production of adenosine triphosphate (ATP), a molecule used to store energy in biological systems, by the phosphorylation of adenosine diphosphate (ADP), ADP + phosphate $\rightarrow$ ATP. The standard-state Gibbs free energy change for this process is 29.3 kJ, so this reaction, by itself will have a very small equilibrium constant. However, by enzymatic reactions, it is coupled to the oxidation of glucose, $C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$, with a standard state Gibbs free energy change of -2807.2 kJ, which is so large that it can drive the phosphorylation of many ADP molecules. In fact, the net overall reaction is

$$C_6H_{12}O_6 + 6O_2 + 38\,\text{ADP} + 38\,\text{phosphate} \rightarrow 6CO_2 + 6H_2O + 38\,\text{ATP}$$

for which $\Delta G^o = -1756.8$ kJ. There are many other examples in biological systems of complex enzymatic reaction networks resulting in one reaction driving another.

SEE ALSO THE FOLLOWING ARTICLES

BIOENERGETICS • HEAT TRANSFER • INTERNAL COMBUSTION ENGINES • PHYSICAL CHEMISTRY • STEAM TABLES

BIBLIOGRAPHY

Pitzer, K. S. (1995). "Thermodynamics," 3rd ed., McGraw-Hill, New York.

Prausnitz, J. M., Lichtenthaler, R. N., and de Azevedo, E. G. (1999). "Molecular Thermodynamics of Fluid-Phase Equilibria," 3rd ed., Prentice-Hall, Englewood Cliffs, NJ.

Rowlinson, J. S., and Swinton, F. L. (1982). "Liquids and Liquid Mixtures," 3rd ed., Butterworths, London.

Sandler, S. I. (1999). "Chemical and Engineering Thermodynamics," 3rd ed., Wiley, New York.

Smith, J. M., Van Ness, H. C., and Abbott, M. M. (1996). "Introduction to Chemical Engineering Thermodynamics," 5th ed., McGraw-Hill, New York.



Thermometry

C. A. Swenson
Iowa State University

T. J. Quinn
Bureau International de Poids et Mesures

I. Introduction
II. Standards and Calibrations
III. Thermodynamic Temperatures
IV. Practical Thermometry

GLOSSARY

Fixed point Unique temperature that is associated with a well-defined thermodynamic state of a pure substance, and that generally involves two or three phases in equilibrium.

Ideal gas Assembly of noninteracting particles. Helium gas at a low pressure is a good approximation for an ideal gas.

International Temperature Scale of 1990 Internationally adopted temperature scale (abbreviated ITS-90 or T_{90}) that provides a reference for all current thermometry.

Primary thermometer Device that directly determines thermodynamic temperatures.

Secondary thermometer Instrument that is used for practical thermometry and that must be calibrated in terms of a primary thermometer.

Standard platinum resistance thermometer Carefully specified secondary thermometer that is used in the definition of the IPTS-68 over much of its range.

Thermodynamic temperature Parameter (actually, an energy) that appears in theoretical calculations of thermal effects.

MODERN THERMOMETRY extends over at least 10 decades in temperature, from the temperatures reached in nuclear cooling experiments to those achieved in nuclear explosions. At both the lowest and the highest extremes, temperatures are measured using methods that are related directly to theory and, hence, correspond to thermodynamic temperatures. At intermediate temperatures, where high accuracy is most necessary, temperatures are defined in terms of secondary thermometers (such as the standard platinum resistance thermometer) that have proved to be stable and sensitive and to have calibrations that vary smoothly with thermodynamic temperature. These instruments serve as interpolation devices between a sequence of accurately defined fixed points to which temperatures have been assigned which correspond closely to thermodynamic values. The thermometers that are used in practical situations may be more convenient to use than either thermodynamic thermometers or scale-defining secondary thermometers, may be smaller

in size, and/or may be more sensitive, while lacking the smoothness and/or stability criteria.

I. INTRODUCTION

The qualitative aspects of temperature and temperature differences are synonymous with the physiological sensations of "hot" and "cold." These descriptions are ambiguous, since often it is the heat conductance or even the thermal mass of the material that is sensed, rather than its actual temperature. Hence, the temperature of a glass object always will seem to be less extreme than that of a metal object, even though the two objects are at the same temperature.

The measurement of temperature, or the science of thermometry, is made quantitative through the observation that the physical properties of materials (density, electrical resistance, and color, for instance) change reproducibly as they become "hotter" or "colder." These changes, which can be relatively large and extremely reproducible for certain well-characterized materials, allow the design and construction of practical thermometers. An important requirement in any science is that measurements made in different localities and in different ways can be related quantitatively, so an agreement on the use of standards must exist. Thermometry standards are based on the observation that certain phenomena always occur at the same, highly reproducible, temperature. The temperatures at which water freezes and then boils under a pressure of 1 atm were recognized very early as being useful thermometric "fixed points," and the Celsius (formerly called centigrade) temperature scale, t, was based on the assignment of 0 and 100°C, respectively, to these two phenomena. As described below, a number of fixed points are used today to define the currently accepted temperature scale.

Once fixed-point temperatures have been assigned, values are associated with intermediate temperatures by interpolation using a "thermometric parameter" that has been evaluated at both lower- and higher-temperature fixed points. This parameter could be, for instance, the expansion of a liquid in a glass bulb (the liquid-in-glass thermometer) or the electrical resistance of a platinum wire (the platinum resistance thermometer; PRT). Since these interpolations may give answers that depend on the material and/or the physical property involved, the standard temperature scale also must designate the type of interpolation device that is to be used. A carefully specified standard platinum resistance thermometer (SPRT) is the designated interpolation instrument over much of the intermediate temperature range, with other instruments important at the extremes of very high and very low temperatures.

The above discussion places no restrictions on what could be an arbitrary assignment of values to the various fixed points, although a "smooth" relationship between these and, for instance, the resistance of an SPRT would appear to be desirable. The concept of a characteristic thermal energy, or of a theoretical temperature, appears both in the science of thermodynamics and in theoretical calculations of thermal properties of materials. Hence, a natural additional requirement is that fixed-point temperatures (and interpolated values) coincide as closely as possible with theoretical (or thermodynamic, or absolute) temperatures, T, which will be measured in kelvins (K). This requirement can be satisfied using a "primary" thermometer, which is a practical device that can be understood completely in a theoretical sense (a gas thermometer, for instance) and that can be used experimentally to study fixed points and interpolation devices. In addition, for purely practical reasons, temperature intervals measured in kelvins and degrees centigrade should have identical numerical values. This was accomplished historically by making measurements with the primary thermometer at the two defining fixed points for the Celsius scale and by requiring that the corresponding temperature difference be exactly 100 K.

Temperatures on the Celsius scale may have either positive or negative values, since 0°C has been chosen arbitrarily, while T must always be positive, except for unusual situations, and $T = 0$ (absolute zero) has a definite meaning (see below). Once the above interval equivalence has been established, t and T will differ by an additive constant, which is the absolute temperature (in K) of the ice point. The triple point of water is much more reproducible than the ice point (see below), and the temperatures of this fixed point are defined to be 273.16 K and 0.01°C. This definition, which establishes the size of the kelvin, was based on the best data available in 1960 for the freezing and boiling points of water on the ideal gas scale. Modern measurements (see below) show that a discrepancy exists between this definition and the definition of the Celsius scale, since the temperature interval between the water freezing and the water boiling points is 99.974 K.

Standards decisions are made by the 48-nation Geneva Conference on Weights and Measures (CGPM), which meets every 4 years (1991, 1995, 1999, etc.). The CGPM acts on the advice of 18 national technical experts who form the International Committee on Weights and Measures (CIPM). The CIPM, in turn, relies heavily on the bench scientists who make up the various Consultative Committees where the actual expertise is located. Thus, it is the Consultative Committee on Thermometry (CCT) that has primary responsibility for establishing and monitoring thermometry standards through recommendations that eventually are acted upon by the CGPM. The work

of the consultative committees is coordinated by the International Bureau of Weights and Measures in Sèvres, just outside Paris, France. The CCT conducts its quality-control role through exchanges of personnel and devices among laboratories and carries out carefully organized international comparisons of thermometer and fixed points. It publishes the results of these exchanges as well as the results of critical evaluations of data. The CCT was responsible for the establishment, in January 1990, of the International Temperature Scale of 1990 (ITS-90), which replaced the International Practical Temperature Scale of 1968 (IPTS-68). Standards decisions are made with great care and after much deliberation, since mistakes have a long lifetime, with, historically, changes being made only every 20 years or so.

II. STANDARDS AND CALIBRATIONS

A. Fixed Points

A useful thermometric fixed point must be reproducible from sample to sample and must exhibit a sharp, well-defined "signal" to which other measurements can be referred easily. In practice, most fixed points are associated with the properties of high-purity, single-component materials. The practical realization of a fixed point with a high accuracy requires considerable care and experience in both the setting-up and the use of the devise, and this is primarily a task for a standards laboratory. Fixed points of all kinds play such an important role in thermometry, however, that they must be a part of a discussion of temperature.

1. Triple Points

The triple point is the unique combination of temperature and pressure at which the liquid, solid, and vapor phases of a pure, single-component system coexist. The triple point of water provides an excellent; illustration of this phenomenon; Fig. 1 is a photograph of a water triple-point cell that is used to realize 273.16 K with an accuracy of 10 μK (10^{-5} K). The glass container contains only pure water, with all traces of air removed. The thermometer is inserted into the central well, around which ice is carefully frozen in a mantle, after which a narrow annulus of water is formed around this well by melting ice from the inside out. Thus, the temperature is uniquely defined since all three phases of pure water are present in equilibrium. The cell in Fig. 1 was removed from its refrigeration chamber for the photograph, but the ring of ice is present, and the thin sheath of water around the well is clearly visible.

Triple points also are important at low temperatures. These are obtained by liquefying a gas (oxygen, argon, neon, and hydrogen are examples) in a sealed system and then carefully cooling it until the solid begins to form at the triple point. Impurities in the starting material can cause changes in the triple-point temperature as the sample is frozen (or melted), and the inherent accuracy of the system (a unique definition of the temperature) is lost. Problems of contamination during gas handling are minimized with a system (Fig. 2) in which a high-purity gas at room temperature and 100 atm is sealed permanently into a carefully cleaned stainless-steel container. As this cell is cooled to the triple point, solid and liquid collect around the copper thermometer well, and the temperature can remain extremely constant as the solid is frozen and then melted. Although these cells have been in use only since 1975, they appear to be remarkably stable with time. The development of sealed triple-point cells (some of which contain several different gases in different parts of the cell) has revolutionized the ease with which low-temperature fixed points can be realized. Similar systems also have been used to obtain high-quality triple points at higher temperatures for other pure materials, with mercury, gallium, and indium metals providing examples.

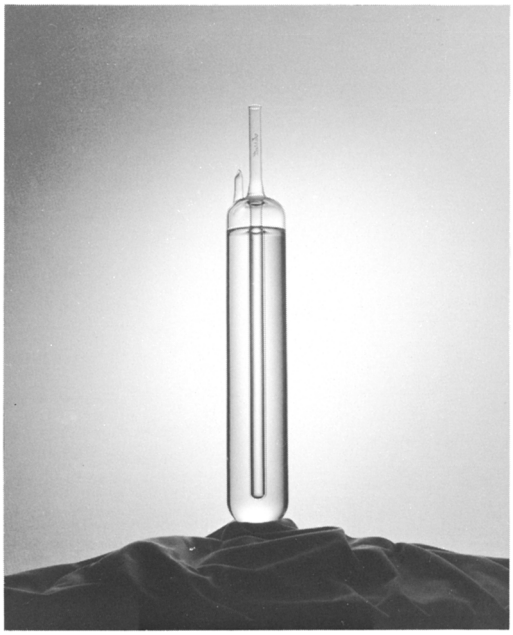


FIGURE 1 A water triple-point cell for use with PRTs. [Courtesy of Jarrett Instrument Company.]

2. Freezing Points

The freezing point is the temperature at which the solid begins to form from the liquid in the presence of atmospheric pressure. The freezing point of water (which defines 0°C), for instance, is approximately 0.01°C lower than the triple point, primarily because the melting temperature of water

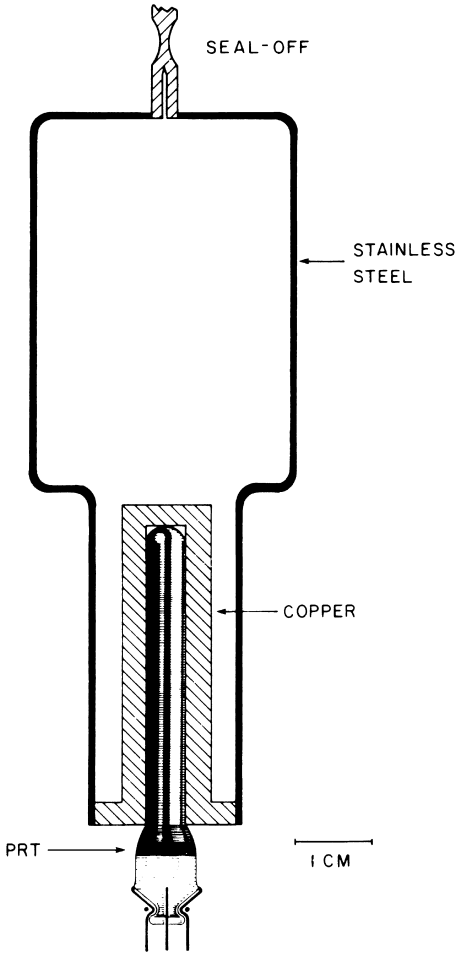


FIGURE 2 An example of the design for a sealed triple-point cell.

is depressed by the application of pressure, although it also is affected by dissolved gases and other impurities. The uncontrollable impurity effects make the freezing point of water less satisfactory as a fixed point than the triple point. To prevent ambiguities, standards thermometry is referred exclusively to the triple point of water, which is defined to be exactly -0.01°C. Melting temperatures generally increase with applied pressure, so the freezing points for most materials are higher than the triple points. Since metals tend to oxidize at high temperatures when exposed to air, atmospheric pressure may be transmitted by an inert gas, but the effect is the same. Again, as for triple points, impurities can destroy the sharpness with which the freezing point can be defined.

3. Boiling Points: Vapor Pressures

The vapor pressure of a pure substance is a unique function of the temperature, so pressure control is equivalent to temperature control. The normal boiling points of pure substances (where the vapor pressure is 1 standard atm, or 101,325 Pa) have been used as fixed points, primarily those of water, oxygen, and hydrogen. Where possible, boiling points have been replaced as fixed points by triple points of other substances to eliminate problems due to pressure measurement and the existence of temperature gradients in the liquid. The vapor pressure–temperature relations for the liquefied helium isotopes, however, often are used directly for the calibration of other thermometers at temperatures from below 1 to 4.2 K. Reliable experimental results for the vapor pressure–temperature relation are available both for the common isotope of mass 4 (^{4}He) and for the much rarer isotope of mass 3 (^{3}He), and equations describing these form the lower temperature portion of the ITS-90. Other vapor pressure–temperature relations (hydrogen, neon, oxygen, nitrogen, oxygen) are useful as secondary standards. In this type of measurement, care must be taken to avoid temperature gradients in the liquid (a sensing bulb is preferred) and cold spots along the pressure measuring tube.

4. Superconducting Transitions

The low-temperature electrical resistance of a number of pure metals disappears abruptly at a well-defined temperature that is characteristic of the metal. These superconducting transition temperatures (T_c) have been developed by the National Institute of Standards and Technology as thermometric fixed points for temperatures from 15 mK (tungsten) to 7.2 K (lead). Early data for polycrystalline materials showed appreciable widths for the transitions, and a corresponding lack of accuracy. Later work on single crystals gives much sharper transitions. The magnitude of T_c depends on the presence of a magnetic field, so care must be taken with magnetic shielding and, also, with the magnitude of the measuring field for the noncontact mutual inductance detection method used to determine T_c.

B. Interpolation Devices

A practical interpolation device must be sensitive, capable of a high accuracy and reproducibility, and convenient to use in different environments. The temperature dependence of its thermometric parameter must be "reasonable," and understood at least qualitatively in a theoretical sense. A very carefully specified form of the platinum resistance thermometer (the SPRT) traditionally has been the interpolation instrument for international scales, and this instrument is used in the definition of the ITS-90 for temperatures from the triple point of hydrogen, 13.8033 K, to the freezing point of silver, 961.78°C. Platinum has the advantages that it can be obtained with a high purity, can be formed easily into wire, has a very high melting point,

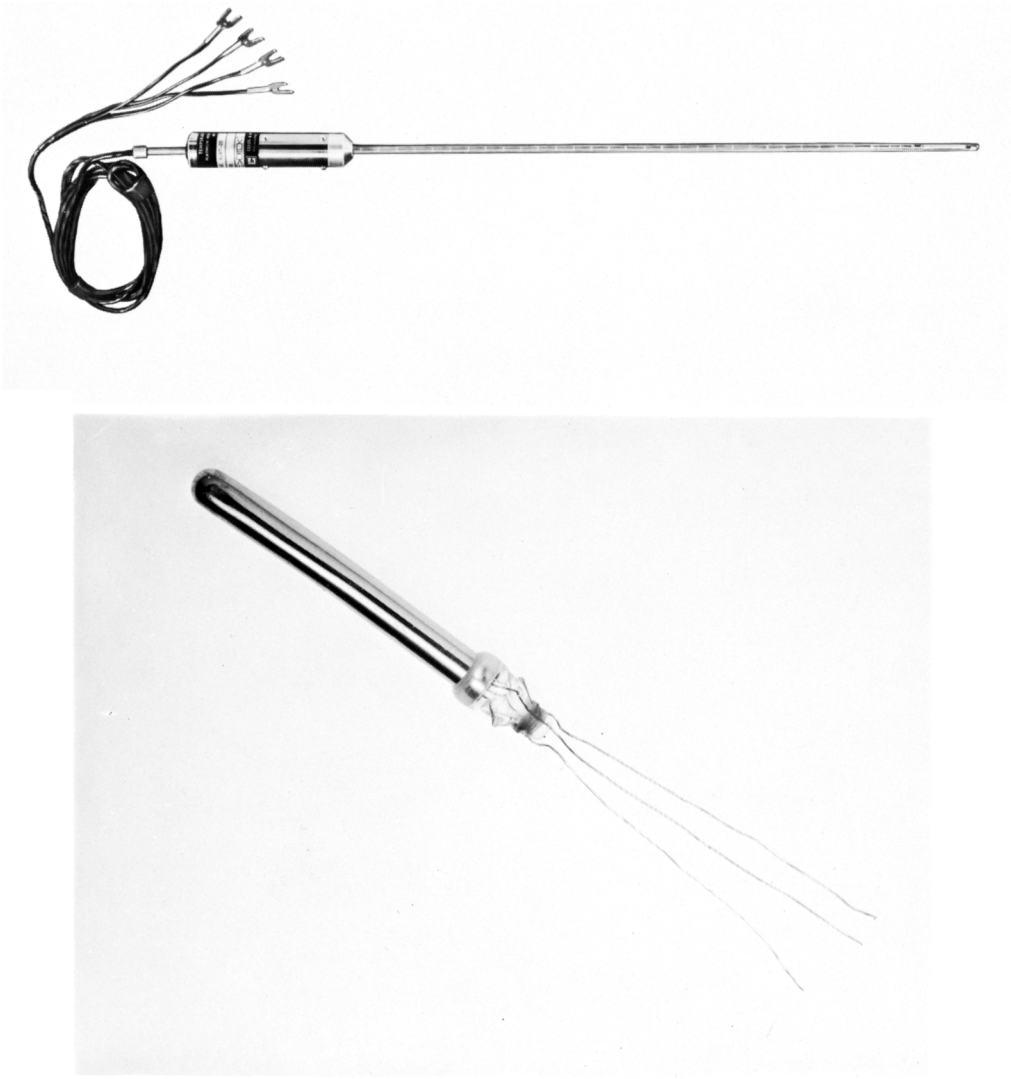


FIGURE 3 Typical standard platinum resistance thermometers. [Courtesy of Yellow Springs Instrument Company.]

and suffers little from oxidation. Many years of use have made the PRT a well-understood instrument both empirically and scientifically.

Figure 3 shows two forms of a commercially available SPRT. In each case, the fine-wire sensing element (typically 25 Ω at the triple point of water) is mounted inside a thin, roughly 6-mm-diameter, 40-mm-long platinum sheath, with a glass or fused quartz seal for introducing the electrical leads. A small amount of "air" provides thermal conductance. A four-lead design allows an unambiguous definition of the resistance of the element. The "capsule" version is intended for low-temperature use, where it can be placed in a vacuum-insulated thermometer well, as for the sealed triple-point cell of Fig. 2. The disadvantage of the capsule form is that the four leads from the resistance element are at the same temperature as the capsule, so leakage resistances between the leads can become important at temperatures greater than 200 or 300°C. The "long-stem" SPRT (Fig. 3, top) reduces this problem since the four leads leave the sealed enclosure at room temperature. Its length, however, makes this instrument impractical for use at temperatures below about 50 K. Internal electrical leakage, which even here becomes a problem for the highest temperatures (above 500°C), can be minimized through the use of long-stem thermometers with ice-point resistances as low as 0.25 Ω. The stability of an SPRT can be determined through periodic checks of its resistance when it is immersed in a triple-point cell (Fig. 1). A good SPRT will give results that are reproducible to better than 0.1 mK even when different triple-point cells are used. The resistance-temperature characteristics of PRTs are discussed specifically in Section IV.

The SPRT becomes relatively insensitive at temperatures below roughly 13.8 K, and the low-temperature calibration is very sensitive to strains that are caused by shock. Other resistance thermometers are more satisfactory for use below 13.8 K (or even 20 K), most importantly those using a rhodium–iron alloy. At the lowest temperatures, the susceptibilities of elementary magnetic systems (electronic to a few millikelvins, then nuclear) show a particularly simple temperature dependence (the Curie–Weiss law; see below) and are used both for interpolation and extrapolation. The melting curve of the helium isotope of mass 3 (^{3}He) also has a strong pressure–temperature relationship below 0.5 K and is being adopted for use as a thermometer for use down to 0.9 mK (see below).

At very high temperatures, above roughly 1000°C, the radiation emitted by a black body can be measured accurately and is used as a measure of temperature (optical pyrometry). Only a single calibration point is required for these measurements, and overlap with the PRT scales is achieved, at least in laboratory measurements. The relative intensities of lines in optical emission or absorption spectra can change with temperature as higher energy levels are excited thermally. These relative intensities can be interpreted directly in terms of T.

C. THE ITS-90

1. The Scale Definition

The currently accepted International Temperature Scale of 1990 differs appreciably from its immediate predecessor (the IPTS-68), with the magnitudes of the differences between the two scales shown in Fig. 4. The lower end of the scale now is 0.65 K rather than 13.8 K, differences from thermodynamic temperatures (especially at low temperature) are reduced to give increased smoothness, and the development of high-temperature SPRTs allows their use to the freezing point of silver (961.78°C). The discontinuity in slope at 630°C in Fig. 4 is related to the change at this temperature in the interpolation instrument which is used to define the IPTS-68. The relatively accurate and precise SPRT was used at lower temperatures, while the much less precise and stable (±0.2 K) platinum–10% rhodium/platinum thermocouple was used to the gold point.

The ITS-90 is defined in terms of the 17 fixed points in Table I, with vapor pressure–temperature relations for the helium isotopes extending the scale definition to 0.65 K. These fixed points are characterized as vapor pressure (v), triple point (tp), or freezing point (fp), with no boiling points being used. The triple point of water is assigned the exact value 273.16 K, with the relationship between the Kelvin and the Celsius temperatures defined as

$$t_{90}/^\circ\mathrm{C} = T_{90}/\mathrm{K} - 273.15; \qquad (1)$$

273.15 appears here instead of 273.16 since, as discussed in the Introduction (Section I), Celsius temperatures are based on the freezing, not the triple, point of water.

The ITS-90 is described most readily in terms of the four interpolation methods (instruments) which are used to define it in four distinct but overlapping temperature ranges. These overlaps represent a change in philosophy

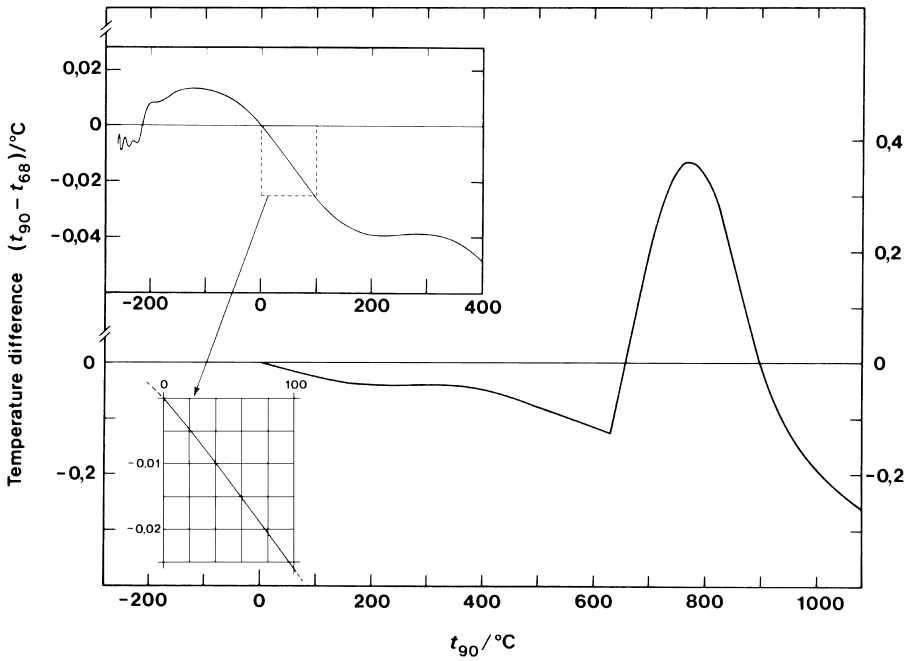


FIGURE 4 Differences between the ITS-90 and its predecessor, the IPTS-68. [From the BIPM.]

TABLE I Fixed-Point Temperatures for the ITS-90

	T_{90} (K)	t_{90} (°C)
1. Helium (v)	3 to 5	−270.15 to −268.15
2. e-Hydrogen (tp)	13.8033	−259.3467
3. e-Hydrogen (v or g)	≈17	≈−256.15
4. e-Hydrogen (v or g)	≈20.3	≈−252.85
5. Neon (tp)	24.5561	−248.5939
6. Oxygen (tp)	54.3584	−218.7916
7. Argon (tp)	83.8058	−189.3442
8. Mercury (tp)	234.3156	−38.8344
9. Water (tp)	273.16	0.01
10. Gallium (fp)	302.9146	29.7646
11. Indium (fp)	429.7485	156.5985
12. Tin (fp)	505.078	231.928
13. Zinc (fp)	692.677	419.527
14. Aluminum (fp)	933.473	660.323
15. Silver (fp)	1234.93	961.78
16. Gold (fp)	1337.33	1064.18
17. Copper (fp)	1357.77	1084.62

from the IPTS-68, since no overlap was allowed between the four ranges which defined that scale.

The low-temperature portion of the ITS-90 is divided into two regions. For the lowest temperatures (0.65 to 5 K), explicit equations are given for the vapor pressure–temperature relations for the two helium isotopes. Temperatures between 3 K and the triple point of neon (24.5561 K) are defined by an interpolating constant volume gas thermometer (see Section III.B.1), which uses either ^{4}He or ^{3}He as the working substance. A procedure is given for correcting the gas thermometer pressures (slightly) for the nonideal behavior of these gases, after which the parameters for a parabolic pressure–temperature relation are determined from the corrected pressures at fixed points 1, 2, and 5 in Table I.

The platinum resistance thermometer (an SPRT) is used to define the ITS-90 from 13.8 K (2 in Table I) to 961.78°C (the freezing point of silver; 15), with the acknowledgment that no single instrument is likely to be usable over this whole range. The characteristics of a real thermometer were used to generate an SPRT interpolation relation which, to obtain the required accuracy, is quite complex. To eliminate differences between thermometers due to different resistances, the primary variable which is used for interpolation is the dimensionless ratio of the thermometer resistance at a given temperature to its value at the triple point of water, 273.16 K,

$$W(T_{90}) = R(T_{90})/R(273.16\ \text{K}). \qquad (2)$$

The triple-point value of R typically is approximately 25 Ω for an SPRT, which will be used from the lowest temperatures to, possibly, 400°C, with smaller values (as low as 0.25 Ω) used for the highest-temperature applications.

A PRT that is acceptable for representing the ITS-90 (an SPRT) must have a high-purity, strain-free platinum element; the ITS-90 defines such an element as one for which either $W(29.7646°\text{C}) \geq 1.11807$ (the gallium triple point) or $W(-38.8344°\text{C}) \leq 0.844235$ (the mercury triple point). An SPRT that is to be used to the freezing point of silver in addition must have $W(961.78°\text{C}) \geq 4.2844$. These requirements eliminate many relatively inexpensive commercial thermometers. A practical requirement which is not stated in the scale is that an SPRT must have a reproducibility at the triple point of water after temperature cycling of better than 1 mK (preferably 0.1 mK). Thermometers which are used above the zinc point (431°C) require careful treatment because of effects due to annealing of the platinum element.

The mathematical functions that are required to describe mathematically the ITS reference interpolation relation for an SPRT are quite complex. For temperatures from 13.8 to 273.16 K, a 13-term power series is required to give $\ln[W_r(T_{90})]$ as a function of $\ln[T_{90}/273.16\ \text{K}]$, while the inverse relation, which gives T_{90} as a function of $W_r(T_{90})$, requires a 16-term power series. The corresponding power series for temperatures from 0 to 961.78°C each contain "only" 10 terms.

Only rarely will the temperature dependence of the resistance for a real thermometer, $W(T_{90})$, agree with that given by the reference function, $W_r(T_{90})$. The values of W and W_r are compared at the various fixed points, and the differences are used to determine the parameters in a deviation function which then is used together with the reference relation to obtain T_{90}. The details again are complex; an SPRT which is to be used from 13.8 to 273.16 K must be calibrated at points 2 through 9 (Table I) to determine the eight parameters in the deviation function. For a calibration which is to be used only within ±30°C of the ice point, the thermometer need only be calibrated at the mercury point, the water triple point, and the gallium point to determine two parameters for the deviation function. All in all, 11 possible subranges are defined; 4 depend on the lowest temperature below 273.16 K at which the thermometer will be used, 1 is for temperatures near 0°C, and 6 depend on the maximum temperature above 0°C at which the thermometer will be used.

A question immediately arises as to the agreement that can be expected between temperatures obtained at, for instance, −15°C, for a given thermometer which has been calibrated using five different procedures and five different sets of fixed points. This is the "uniqueness" problem. The belief is that the differences at a given temperature between calibrations using different ranges will be comparable with differences between different thermometers which are calibrated in a given range. This "nonuniqueness" will

be a few tenths of a millikelvin near room temperature, less than 1 mK for the more extreme parts of the scale between 13.8 K and 420°C, and should be less than 5 mK at the highest temperatures.

The highest range of the ITS-90, above the silver point, is defined by optical pyrometry, using Planck's law to obtain the radiant emission from a black-body cavity for a given wavelength, λ, and bandwidth. The ratio of the spectral radiances at the temperature T_{90} and at the reference temperature, X, is related to the absolute temperature by

$$\frac{L_\lambda(T_{90})}{L_\lambda(T_X)} = \frac{\exp[c_2/\lambda T_{90}(X)] - 1}{\exp[c_2/\lambda T_{90}] - 1}, \qquad (3)$$

where $T_{90}(X)$ refers to any one of the silver [$T_{90}(\mathrm{Ag}) = 1{,}234.93$ K], the gold [$T_{90}(\mathrm{Au}) = 1{,}337.33$ K], or the copper [$T_{90}(\mathrm{Cu}) = 1357.77$ K] freezing points. Here the optical pyrometer both defines the scale and serves as the interpolation device. The ITS-90 specifies the use of the theoretical value for the constant c_2, so there are no adjustable parameters in this relation. Proper realization of temperatures by pyrometry requires care in the design of the cavities in which the gold and the sample are located, and as with most thermometry, care must be taken to avoid systematic errors.

2. Calibration Procedures

Working thermometers (either transfer standards or working instruments) should be calibrated by following the procedures outlined in the basic ITS-90 document to reproduce the scale. In practice, this can be a cumbersome procedure, especially at low temperatures, where gas thermometry requires long-term experiments. In this temperature region, gas thermometry results will be transferred to highly stable rhodium–iron resistance thermometers, and most subsequent calibrations will be carried out in terms of "point-by-point" comparisons at thermal equilibrium between a set of standard thermometers and the unknown thermometer(s). This also may be true for higher, PRT, temperatures when calibrations are not carried out at a national standards laboratory. In this instance, "standards" which have been calibrated directly on the ITS-90 may be used as substitutes for true fixed point devices. Three standard thermometers are the useful minimum, since not more than one would be expected to show drift (instability) in any given period of time. The result is a table of temperatures and corresponding W's, with the W's converted to $R(T_{90})$ using the measured $R(273.16\ \mathrm{K}) = R_o$ to eliminate dependence on a standard resistance value. To a first approximation, small changes in R_o will have little effect on the $W(T_{90})$ relationship for a thermometer.

For moderate and low temperatures, the sheaths of the thermometers can be inserted in individual mounting holes in an isothermal metal block. Thermal shielding of the block, anchoring of the leads to the block, vacuum insulation, and temperature control all are important factors in such a thermometer comparator. Variable-temperature baths (oil or possibly molten salt) are used at higher temperatures where long-stem thermometers must be used. Calibrations carried out by each of the national standards laboratories can be expected to be equivalent, and to represent the ITS-90 within stated uncertainties. Other calibration sources, which generally are traceable to a national standards laboratory, generally have less rigorous controls, and care must be taken in assessing the accuracy of calibrations that are supplied. If accuracy is important, the performance of a thermometer can be spot-checked with commercially available sealed fixed-point devices, with gallium (see Table I) being most useful near room temperature. This may be particularly important when highly accurate thermometry is required for the maintenance of standards or for biological studies.

D. Electrical Measurements

High-quality electrical measurements traditionally have used very accurate dc techniques. Voltages were measured potentiometrically in terms of standard cells, while resistances were measured using Wheatstone or other types of bridges. For accurate work, a standard resistor or a resistance thermometer is designed with four terminals, two of which are for the measuring current, while the second pair, mounted just inside the current leads at each end, measures the potential drop across the resistor. If a conventional Wheatstone-type bridge technique is used, the bridge determines the sum of the resistances of the resistor and of the leads, so a separate measurement of the resistance of a pair of leads at one end of the resistor (or thermometer) must be made. Care must be taken that the lead resistances are symmetrical. These measurements can be simplified if a potentiometer is used to compare directly the potential drops across a standard resistor and the unknown for a common current. In this case, negligible current flows through the potential leads, and no lead correction is required.

In both bridge and potentiometric measurements, parasitic emfs (voltages) can exist in the lead wires and the measuring instrument, with current reversal required to eliminate their effects. In addition, since the bridge contains standard resistances of various magnitudes, these must be intercompared and recalibrated regularly to detect aging effects. The linearity of a dc potentiometer also must be calibrated at regular intervals for the same reason.

Modern semiconductor technology has caused major changes in the above procedures. First, voltmeters now routinely have extremely high input impedances

(greater than 1000 MΩ) and linearities at the 10^{-6} level. Hence, most accurate electrical measurements now are made using these instruments rather than potentiometers or bridges. Modern multimeters often can be used in a four-terminal mode for resistance measurement, and most can be interfaced directly with a computer for experimental control and data acquisition.

When the highest accuracy in resistance measurement is required, variations of the potentiometer technique are used in which the accurate division of voltage levels is carried out using ratio transformers rather than resistive windings. These components are very similar to ideal transformers or inductors, with windings on a high-permeability mumetal toroid system for which the stability is determined by winding geometry rather than a physical property. The current comparator is a dc instrument in which the condition for zero magnetic flux in a core is used to determine the ratio of currents through two resistances (a standard and an unknown) when the potential drops across them are equal. The effects of parasitic voltages are eliminated by using current reversal. These instruments are in common use in standards laboratories and are capable of determining resistance ratios potentiometrically at the 10^{-8} level. This corresponds to better than 10 μK for an SPRT with a 25-Ω ice-point resistance and is better than the long-term stability of many standard resistances. It is for this reason that SPRT measurements are always expressed in terms of Eq. (2), using a direct determination of $R(273.16\ \mathrm{K})$.

Various alternating current bridges and potentiometers have been constructed using ratio transformer techniques. Figure 5 shows a very simple version of an ac ratio-transformer bridge. The ac voltage drop across an unknown resistor is compared with a fraction of the voltage drop across a standard resistor. This fraction, which is determined by the turns ratio, is adjusted until a null is indicated at the detector. Typically, this is a phase-sensitive detector with transformer input and a sensitivity to extremely low (nV; 10^{-9} V) voltages. This bridge is useful primarily for temperature control, since the finite input impedance of the transformer (typically 10^5 Ω at 400 Hz) causes unacceptable shunting of the reference resistor. The input impedance of the transformer can be increased greatly by sophisticated designs that use multiple cores and windings and operational amplifier feedback. As a result, accuracies of 10^{-8} are also reported for the ac measurement of a standard 25-Ω SPRT.

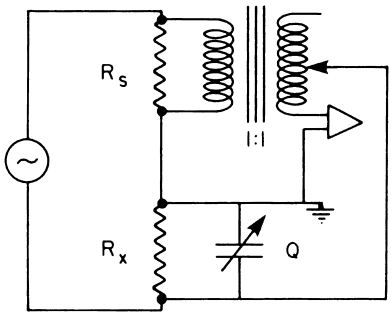


FIGURE 5 An elementary ac ratio-transformer bridge for resistance measurements.

Although the effects of parasitic dc voltages are eliminated with ac methods, frequency-dependent lead admittance effects (due to shunt capacitances between thermometer leads) are important, and both in-phase and quadrature balance conditions must be met. This is accomplished in Fig. 5 with the variable shunt capacitor. It is for this reason that ac bridges are restricted to relatively low resistance values for the most accurate work.

III. THERMODYNAMIC TEMPERATURES

A. General Concepts

The concept of thermodynamic temperature arises from the second law of thermodynamics and the existence of reversible heat effects, such as for the isothermal compression of an ideal gas. The maximum (Carnot) efficiency for a heat engine, for example, is expressed in terms of a ratio of thermodynamic temperatures.

Developments of statistical mechanics contain a characteristic energy that is the same for all systems that are in thermal equilibrium and that increases as the internal energy of a system is increased. This characteristic energy has properties that are identical to those of temperature as it is defined in both the thermodynamic and the practical senses. This characteristic energy appears in an elementary manner in the Boltzmann factor, which determines the relative populations of two states that are separated by an energy difference ΔE,

$$N_1/N_2 = \exp(-\Delta E/k_{\mathrm{B}}T). \tag{4}$$

In this expression, $k_{\mathrm{B}}T$ is the characteristic energy, and k_{B} (as yet undetermined) is the Boltzmann constant. Equation (4) suggests that the concept of a level of temperature is purely relative. A collection of systems can be said to be at a low temperature (close to $T = 0$) if most (all) of them are in their lowest energy (ground) state, that is, if $\Delta E \gg k_{\mathrm{B}}T$. Alternatively, a high temperature corresponds to an equal population of the levels. Whether or not a temperature is "high" or "low" thus depends on the characteristic energies of the system and is a purely relative concept. Absolute zero corresponds to a state at which every conceivable system is in its ground state. Negative temperatures occur when (as in some laser systems) an

upper metastable level has been forced to have a larger population than a lower level.

The relationship between theoretical and practical temperatures (see Section I) has been determined most often using measurements made with an ideal gas. The experimental equation of state for such a system is written

$$PV_m = RT, \tag{5}$$

with V_m the volume per gram molecular weight of the gas, R the gas constant per mole (8.317 J/F mol-K), and T related to the Celsius scale by Eq. (1). Since a Carnot heat engine with an ideal gas as the working medium has an efficiency identical to that of a Carnot cycle, T as it appears in Eq. (5) can be chosen to be equal to thermodynamic temperatures.

Statistical mechanics as applied to an ideal gas (a collection of noninteracting particles) also gives Eq. (5), if RT is assumed to be proportional to the characteristic thermal energy of the system and to the total number of particles. The association with Eq. (4) exists through the introduction of the gas constant per molecule, the Boltzmann constant, $k_B = R/N_A$, where N_A, the Avagadro constant, is the number of molecules in a gram molecular weight of a substance. The characteristic thermal energy that appears in the Boltzmann relation is the same as that which appears in the ideal-gas law.

B. Absolute or Primary Thermometers

The use of fixed points and designated interpolation instruments would not be necessary if an absolute or primary thermometer could be used directly as a practical thermometer. A single calibration of such a thermometer at the triple point of water (273.16 K) would serve to standardize the thermometer once and for all. Unfortunately, most primary thermometers are relatively clumsy devices and may require elaborate instrumentation and possibly long equilibrium and/or measurement times.

Two exceptions are the optical pyrometer at high temperatures and the magnetic thermometer at low temperatures. In each of these cases, data are taken using the primary thermometric parameter, with this parameter related directly by theory to the absolute temperature. At intermediate temperatures, fixed points and easily used secondary thermometers must be used for the routine measurement of temperature. Primary thermometers, then, are used to establish the temperatures that are assigned to the fixed points and to test the smoothness and appropriateness of the calibration relations that are used with the secondary thermometers.

The following sections discuss briefly the various types of primary thermometers that have been used to obtain accurate thermodynamic temperatures. Gas thermometry in various forms traditionally has been of primary importance in this area, but modern optical pyrometry has comparable importance at high temperatures, and noise and magnetic thermometry also have had important complementary roles. The existence of several approaches for a given temperature range is important to provide confidence in the relationship between theory and experiment, and to provide information about the possible existence of systematic errors.

1. Gas Thermometry

The ideal-gas law [Eq. (5)] is valid experimentally for a real gas only in the low-pressure limit, with higher-order terms (the virial coefficients, not defined here) effectively causing R to be both pressure and temperature dependent for most experimental conditions. While these terms can be calculated theoretically, most gas thermometry data are taken for a variety of pressures, and the ideal-gas limit, and, hence, the ideal-gas temperature, is achieved through an extrapolation to $P = 0$. The slope of this extrapolation gives the virial coefficients, which are useful not only for experimental design, but also for comparison with theory. The following discussion of ideal-gas thermometry is concerned, first, with conventional gas thermometry, then with the measurement of sound velocities, and, finally, with the use of capacitance or interferometric techniques. Each of these instruments should give comparable results, although the "virial coefficients" will have different forms.

Gas thermometry in the past 20 years or so has benefited from a number of innovations that have improved the accuracy of the results. Pressures are measured using free piston (dead weight) gauges that are more flexible and easier to use than mercury manometers. The thermometric gas (usually helium) is separated from the pressure-measuring system by a capacitance diaphragm gauge, which gives an accurately defined room-temperature volume and a separation of the pressure-measurement system from the working gas. In addition, residual-gas analyzers can determine when the thermometric volume has been sufficiently degassed to minimize desorption effects.

In isothermal gas thermometry, absolute measurements of the pressure, volume, and quantity of a gas (number of moles) are used with the gas constant to determine the temperature directly from Eq. (5). Data are taken isothermally at several pressures, and the results are extrapolated to $P = 0$ to obtain the ideal-gas temperature as well as the virial coefficients. A measurement at 273.16 K gives the gas constant.

A major problem in isothermal gas thermometry is determining the quantity of gas in the thermometer, since this ultimately requires the accurate measurement of a small difference between two large masses. Most often,

this problem is bypassed by "filling" the thermometer to a known pressure at a standard temperature, with relative quantities of gas for subsequent fillings determined by division at this temperature between volumes that have a known ratio. The standard temperature may involve a fixed point or, for temperatures near the ice point, an SPRT that has been calibrated at the triple point of water. Since the volume of the gas for a given filling is constant for data taken on several subsequent isotherms, and the mass ratios are known very accurately, the absolute quantity of gas needs to be known only approximately. Excellent secondary thermometry is very important to reproduce the isotherm temperatures for subsequent gas thermometer fillings. The results for the isotherms (virial coefficients and temperatures) then are referenced to this standard "filling temperature."

The procedure for constant-volume gas thermometry is very much the same as that for isotherm thermometry, but detailed bulb pressure data are taken as a function of temperature for one (and possibly more) "filling" of the bulb at the standard temperature. To first order, pressure ratios are equal to temperature ratios, with thermodynamic temperatures calculated using known virial coefficients. In practice, the virial coefficients vary slowly with temperature, so a relatively few isotherm determinations can be sufficient to allow the detailed investigation of a secondary thermometer to be carried out using many data points in a constant-volume gas thermometry experiment. If the constant-volume gas thermometer is to be used in an interpolating gas thermometer mode (as for the ITS-90), the major corrections are due to the nonideality of the gas. When a nonideality correction is made using known values for the viral coefficients, the gas thermometer can be calibrated at three fixed points (near 4 and at 13.8 and 24.6 K) to give a quadratic pressure–temperature relation that corresponds to T within roughly 0.1 mK.

The velocity of sound in an ideal gas is given by

$$c^2 = (C_P/C_V)RT/M, \tag{6}$$

where the heat capacity ratio (C_P/C_V) is 5/3 for a monatomic gas such as helium. Since times and lengths can be measured very accurately, the measurement of acoustic velocities by the detection of successive resonances in a cylindrical cavity (varying the length at constant frequency) appears to offer an ideal way to measure temperature. This is not completely correct, however, since boundary (wall and edge) effects that affect the velocity of sound are important even for the simplest case in which only one mode is present in the cavity (frequencies of a few kilohertz). These effects unfortunately become larger as the pressure is reduced. An excellent theory relates the attenuation in the gas to these velocity changes, but the situation is very complex and satisfactory results are possible only with complete attention to detail. An alternative configuration uses a spherical resonator in which the acoustic motion of the gas is perpendicular to the wall, thus eliminating viscosity boundary layer effects. The most reliable recent determination of the gas constant, R, is based on very careful sound velocity measurements in argon as a function of pressure at 273.16 K, using a spherical resonator.

The dielectric constant and index of refraction of an ideal gas also are density dependent through the Clausius–Mossotti equation,

$$(\varepsilon_r - 1)/(\varepsilon_r + 2) = \alpha/V_m = \alpha RT/P, \tag{7}$$

in which $\varepsilon_r (= \varepsilon/\varepsilon_0)$ is the dielectric constant and α is the molar polarizability. Equation (7) suggests that an isothermal measurement of the dielectric constant as a function of pressure should be equivalent to an isothermal gas thermometry experiment, while an experiment at constant pressure is equivalent to a constant-volume gas thermometry experiment. The dielectric constant, which is very close to unity, is most easily determined in terms of the ratio of the capacitance of a stable capacitor that contains gas at the pressure P to its capacitance when evacuated. The results that are obtained when this ratio is measured using a three-terminal ratio transformer bridge are comparable in accuracy with those from conventional gas thermometry. An advantage is that the quantity of gas in the experiment need never be known, although care must be taken in cell design to ensure that the nonnegligible changes in cell dimensions with pressure can be understood in terms of the bulk modulus of the (copper) cell construction material.

At high frequencies (those of visible light), the dielectric constant is equal to the square of the index of refraction of the gas ($\varepsilon_r = n^2$), so an interferometric experiment should also be useful as a primary thermometer. No results for this type of experiment have been reported, however.

2. Black-Body Radiation

The energy radiated from a black body is a function of both temperature and wavelength [Eq. (3)]. An ideal black body has an emissivity (and hence an absorptivity) of unity, or a zero reflectivity. The design of high-temperature black bodies to satisfy this condition requires considerable care. In practice, a usable design would consist of a long cylindrical graphite cavity with a roughened interior that is, for instance, surrounded by freezing gold to maintain isothermal conditions. The practical aspects of optical pyrometry are discussed briefly in Section IV. For the present purposes, optical pyrometry using well-defined wavelengths and sensitive detectors (so-called photon-counting techniques) can be used with Eq. (3) to measure relative

temperatures with a high accuracy (better than 10 mK) at temperatures as low as the zinc point, 419.527°C. This gives a valuable relationship between the high temperature end of current gas thermometry experiments and the temperatures that are assigned to the gold and silver points.

The total energy that is radiated by a black body over all wave lengths [the integrated form of Eq. (3)] is the well-known Stefan–Boltzmann law,

$$dW/dT = \sigma T^4. \tag{8}$$

Here, $\sigma = (2\pi^5 k_B^4/15c^2h^3) = 5.67 \times 10^{-8}$ W/m^{2}K^4 is the Stefan–Boltzmann constant. Measurements of the power radiated from a black body at 273.16 K give σ directly, and, since both Planck's constant, h, and the velocity of light, c, are well known, also give the Boltzmann constant, k_B. Relative emitted powers also give temperature ratios. Total radiation measurements [Eq. (8)] have been carried out for black bodies in the range from −130°C to + 100°C using an absorber at a low temperature (roughly 2 K) to measure the total radiant power that is emitted.

3. Noise Thermometry

Noise thermometry is another, quite different, system that can be understood completely from a theoretical standpoint and that can be realized in practice. The magnitude of the mean-square thermal noise voltage (Johnson or Nyquist noise) that is generated by thermal fluctuations of electrons across a pure electrical resistance, R, is given by

$$(V^2)_{avg} = 4k_B TR\Delta f. \tag{9}$$

This simple exact expression assumes that R is frequency independent, with the mean-square noise voltage depending on R and the bandwidth in hertz, Δf, over which the measurement is made. These measurements are difficult, since, to achieve the needed accuracy, consistent measurements must be made of the long-time average of the square of a voltage. In most instances, the results are obtained as the ratio of the mean square voltage at T to that at a standard temperature (possibly 273.16 K), so the absolute values of the voltages need not be determined. Instrumental stability is very important, however. Noise temperatures have been determined from as low as 17 mK [17×10^{-3} K, using SQUID (Superconducting Quantum Interference Device) technology] to over 1000°C. While noise thermometry is difficult to carry out in a routine fashion, the measurements involved are so different from those for gas thermometry and optical pyrometry that the results are extremely useful.

4. Magnetic Thermometry

The magnetic susceptibility of an ideal paramagnetic salt (a dilute assembly of magnetic moments) obeys Curie's law,

$$x = C/T, \tag{10}$$

where C, the Curie constant, is proportional to the number of ionic magnetic moments and their magnitudes. The magnetic moments may be due either to electronic or to nuclear effects, with a difference in magnitude of roughly 1000. Interactions between the moments eventually cause the breakdown of Eq. (10) at temperatures of the order of millikelvins (or higher) for electronic paramagnetism, and at temperatures 1000 times smaller for nuclear systems.

Magnetic thermometry involving electron spins is not strictly primary thermometry, since the number of moments in the sample cannot be determined with any precision, and Curie's law is obeyed only approximately for any real system. Magnetic interactions between the moments and complications due to the existence of excited states for the ions cause difficulties in almost every case. An ion can be chosen for which the excited states are not populated for a given experiment, with deviations due to magnetic interactions expected on theoretical grounds to give first-order corrections to Curie's law which are of the form

$$x = A + B/(T + \Delta + \delta/T). \tag{11}$$

The parameter A is due to temperature-independent diamagnetism and paramagnetism, while Δ represents effects due to surrounding moments, and δ arises because of complex spin systems. In practice, each of these parameters must be determined empirically.

While a paramagnetic salt such as cerium magnesium nitrate [CMN, $Ce_2Mg_3(NO_3)_{12}\cdot 24H_2O$] shows almost-pure Curie law behavior ($\Delta = 0.3$ mK, $\delta = 0$), the dilution of its moments and consequent small susceptibility make measurements difficult above 2 K, with a breakdown of Eq. (11) arising near 4 K due to the beginning occupation of a higher-energy state of the cerium ion. Even at low temperatures, controversy exists for CMN as to the meaning of the "nonideality" parameters, and the significance of different values of Δ for single-crystal and powdered samples. The use of SQUID technology rather than conventional ratio-transformer mutual inductance bridges allows measurements to be made with extremely small samples. Paramagnetic salts with larger susceptibilities, which are useful at higher temperatures, will have larger values for the nonideality parameters and will show deviations from even Eq. (11) at temperatures not far below 1 K.

5. Helium Melting-Pressure Thermometry

At temperatures below the lower limit of the ITS-90, 0.65 K, a new low-temperature scale is being proposed by the CCT based on the relation between the pressure and the temperature of melting ^{3}He. Although the helium melting temperature–pressure relation used in the new scale is closely related to the Clausius–Clapyron equation its temperature cannot be calculated directly from this equation with sufficient accuracy. Instead, the relation is based on experimental measurements using magnetic thermometry, noise thermometry, and nuclear-orientation thermometry. It is thus not strictly a primary thermometer. The new scale is referred to as to the "Provisional Low-Temperature Scale, 0.9 mK to 1 K: PLTS-2000." The scale is defined by the relation between the temperature of melting ^{3}He and fixed points, i.e., the minimum in the melting pressure of ^{3}He at a temperature of about 315 mK and a pressure of 2.93 MPa and at the A, A–B, and Néel transitions in ^{3}He at temperatures of about 2.44, 1.9, and 3.44 mK respectively.

6. Nuclear Orientation Thermometry

At temperatures below 100 mK or so, the splitting of nuclear energy levels in a single crystal may become comparable with the characteristic thermal energy, $k_B T$. The γ-ray emissions from the oriented nuclei then may be anisotropic, and the anisotropies can be used to determine the relative populations of these levels. In the simplest possible two-level case, Eq. (4) can be applied to obtain the temperature directly from these nuclear orientation experiments. Such measurements have been made from 10 to roughly 50 mK for radioactive cobalt of mass 60 in a single-crystal nonradioactive cobalt lattice. These have confirmed SQUID noise measurements in the assignment of absolute temperatures to the superconducting transitions of the National Bureau of Standards SRM 768 device. The energy levels of the nuclei involved must be understood in detail from other measurements before these methods can be used, but, again, it is useful that two independent measurements can be used to assign thermodynamic temperatures in an extreme region of the temperature spectrum.

7. Spectroscopic Methods

Optical spectroscopy can give information about the relative populations of excited states in a very high-temperature system, such as a plasma. This information then can be combined with the Boltzmann relation or direct theoretical calculations to obtain the temperature directly, as for nuclear orientation experiments. Again, the system must be understood theoretically, and possible complications due to interactions must be recognized. This use of spectroscopic data for primary thermometry represents the only possible means for determining extremely high temperatures.

C. The ITS-90 and Thermodynamic Temperatures

Each of the above primary thermometers has been used for at least a limited temperature region in the establishment of the ITS-90. At the lowest temperatures, the scale is based on a combination of results from magnetic, noise, and gas thermometry, with several gas thermometry experiments of most importance from liquid helium and/or liquid hydrogen temperatures to 0°C. These agree well with total-radiation experiments at temperatures above 240 K. Gas thermometry results overlap pyrometry data for temperatures from 457 to 661°C, and the comparison of an SPRT with pyrometry data provided the SPRT reference function for temperatures from 660°C to the silver point. The correspondence between the ITS-90 and thermodynamic temperatures is believed to vary from ±0.5 mK at the lowest temperatures to a maximum of ±2 mK for any temperature below 0°C. At higher temperatures, the possible difference rises from ±3 mK at the steam point to ±25 mK at 660°C. The three highest temperature reference points (based on freezing points for silver, gold, and copper) are expected to be internally consistent to within the accuracy of standards pyrometry and to have potential differences from thermodynamic temperatures of ±0.04, 0.05, and 0.06 K, respectively, which reflect the uncertainties at the primary reference temperature of 660°C. The most important characteristic of the ITS-90, however, is that it is believed to be smoothly related to T at all temperatures, with no abrupt differences in slope such as appear in Fig. 4, where, on the scale of this figure, T_{90} is identical to T.

IV. PRACTICAL THERMOMETRY

Many types of thermometers are in general use, and many more have been proposed. The following is a brief summary of the characteristics of the more common types of secondary thermometers, with no attempt made to be complete or comprehensive. The choice of a type of thermometer for a given application is somewhat arbitrary, with the deciding factors sometimes dictated by rigorous constraints but more often by personal preferences and/or prejudices. The accuracy or longevity of a thermometer calibration (a certificate or a table) should not be taken for granted when a temperature must be known within specified limits. Checks should be made, either in terms of a close-by fixed point (the freezing point of water and the

triple point of gallium are particularly useful near room temperature) or by comparison with one, but preferably two or more, carefully handled, "standard" thermometer. An electrical instrument should never be relied upon to give answers that are correct to all of the significant figures that are generated in the display or in the printout, especially if important conclusions depend on these numbers.

A. Liquid-in-Glass Thermometers

These represent the oldest, and still very common, practical thermometers, although they are increasingly being replaced by low-cost electronic devices using semiconductor elements (see below) as the temperature sensor. They come in many forms and qualities with a variety of liquids, although mercury is the choice for accurate applications. A very good thermometer for use up to 100°C can be calibrated to 0.01°C or better and will remain stable at this level for a considerable period of time. Care must be taken in the use of such a thermometer, since the readings depend on the depth of immersion of the thermometer. Thus, they are most useful for measurements on liquids where a surface is defined. The disadvantage of liquid-in-glass thermometers is that they must be calibrated manually, a tedious process, and must be read by eye, with no opportunities for automated data acquisition.

B. Resistance Thermometers

Resistance thermometers, or, more strictly, thermometers for which a voltage reading depends on an applied current, quite naturally fall into two categories. The first includes pure metals and metallic alloys that exhibit a positive temperature coefficient of resistance. Alloys with very small coefficients are useful for constructing the standard resistances that must play an important role in the practical use of resistance thermometers. The second category includes primarily semiconducting materials, for which the temperature coefficient of resistance is negative. It also includes devices, such as diodes, for which the forward voltage is a function of temperature.

General considerations for the measurement of electrical resistance, discussed in Section II.D, are not repeated here. The reproducibility of a practical resistance thermometer is an important characteristic that is not always directly related to the cost. Its calibration also may depend critically on the magnitude of the measuring current, so care should be taken to follow the manufacturer's (or calibrator's) recommendations. Resistance thermometers often are used both for the control of temperature (as in a thermostat) and for the measurement of the temperature. In general, this is not a recommended procedure, since a temperature-control sensor generally is located in the vicinity of the source of heat of refrigeration and will not give a true average reading for the volume that is being controlled.

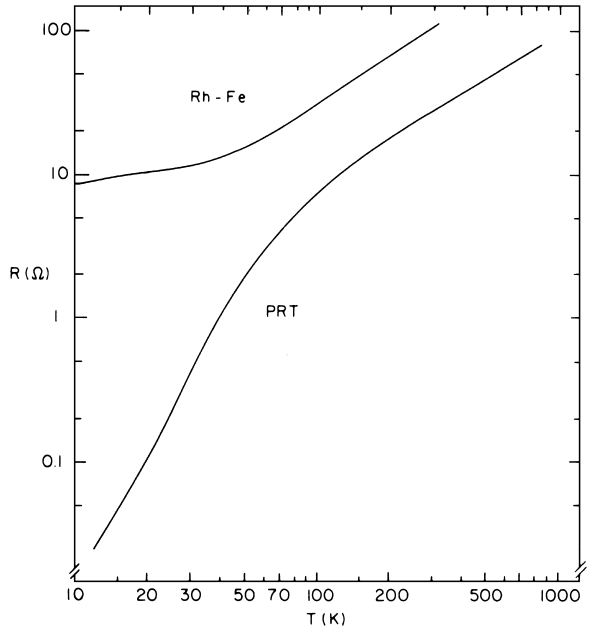


FIGURE 6 The temperature dependences of the resistances for two metallic resistance thermometers.

1. Metallic Thermometers

The platinum resistance thermometer (PRT) is a typical metallic thermometer; the temperature dependence of the resistance that is shown in the double-logarithmic plot in Fig. 6 is characteristic of most metals. Near room temperature and above, the electrical resistance of a pure metal is associated primarily with lattice vibrations and is proportional to T, with the temperature coefficient of resistance approximately independent of temperature. Impurity effects end to dominate at low temperatures, where the resistance approaches a constant value as T approaches zero. The ratio of the room-temperature resistance to its low-temperature value (the resistance ratio) is a measure of the purity of a metal, and the ratio of 1000 for the SPRT in Fig. 6 (the nominal ice point resistance is 25 Ω) is characteristic of a very pure metal.

Industrial PRTs are constructed from a "potted" wire or a thin film bonded to a ceramic substrate. These have a characteristic resistance very similar to that of an SPRT near room temperature but have a relatively high value for the low-temperature resistance due to the quality of the platinum and also to the strains induced in fabrication. Standard calibration tables exist for these commercial PRTs for temperatures from 77 K upward,

with the objective of allowing routine substitution and replacement of thermometers as needed. One of the difficulties in using pure metallic thermometers at temperatures below 20 K is that the resistance is very sensitive to strains that are induced by shocks, so great care must be taken in handling a calibrated SPRT. Hence, a PRT that was not wound in a strain-free configuration could be expected to be relatively more unstable than the much more expensive SPRT. An additional characteristic of inexpensive PRTs is that they are primarily two-lead devices. For most applications, it is useful to attach a second pair of leads so that the resistance of the thermometer is well defined.

The temperature dependence of an alloy thermometer is also shown in Fig. 6. The primary component of this thermometer is rhodium metal, with a slight amount (0.5%) of iron added as an alloying agent. The localized magnetic moment of the iron scatters electrons very well at low temperatures and is responsible for the relatively high 10 K resistance for this thermometer, which has a nominal 100-Ω room-temperature resistance. The interaction of these iron moments with the electrons also results in an approximately linear temperature dependence for the low-temperature resistivity, in contrast with the SPRT, as shown in Fig. 7 for temperatures to 0.25 K. This thermometer is much more satisfactory than the PRT at low temperatures because of both its sensitivity and its stability. The wire is extremely stiff and difficult to fabricate into a thermometer element. As a result, the thermometers are very insensitive to shock, and aging and annealing effects are virtually nonexistent. Rhodium thermometers, which are packaged similarly to SPRTs, now form the basis for most practical low-temperature standards thermometry. They are available also in other packages for use in practical measurements, possibly (as Fig. 6 indicates) for temperatures up to room temperature. A single thermometer that can be used with a reasonable sensitivity from 0.5 to 300 K is a very useful device.

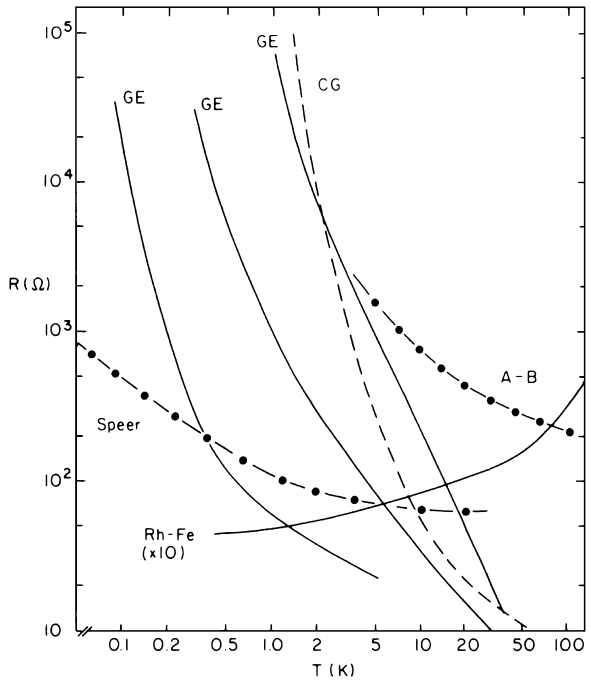


FIGURE 7 The resistance–temperature relations for several low-temperature thermometers. [The GE and CG results are through the courtesy of Lake Shore Cryotronics, Inc.]

2. Semiconductors

Figure 7 gives, along with low-temperature results for a rhodium–iron thermometer, a double-logarithmic plot of the resistance–temperature relationships for a number of low-temperature thermometers which are constructed from semiconducting materials. This presentation does not include an R-vs-T relationship for another often-used semiconducting thermometer, the thermistor (see below), which would be similar to that for the carbon–glass (CG) thermometer, but for higher temperatures.

Commercial radio resistors were used as the first semiconducting low-temperature thermometers, with the most popular being, first, those manufactured by Allen–Bradley (A-B), and, later, those manufactured by Speer. The bonding of the electrical leads to the composite material in these resistors proved to be quite rugged, and although small (occasionally large) resistance shifts occurred on subsequent coolings to liquid helium temperatures, the calibrations remained stable as long as the thermometers were kept cold. The thermometric characteristics of these two brands of resistors have the common feature that the temperature coefficient of the resistance is a smooth and monotonic function of the temperature. The details of their temperature variation are seen to be quite different, however, with the A-B resistors being very sensitive, while the Speer resistors have a reasonable resistance even at the lowest temperatures. These resistors are still used for low-temperature measurements, although improvements in their composition have changed (and downgraded) their thermometry characteristics. The carbon–glass thermometer, which uses fine carbon filaments deposited in a spongy-glass matrix, also has a well-behaved resistance–temperature characteristic, as well as a high sensitivity. This thermometer suffers from lead-attachment problems and has instabilities (minor for many purposes) that make it unsuitable for standards-type measurements. All three of these thermometers have resistances with moderate magneto-resistance characteristics so are useful for measurements in a magnetic field.

Germanium resistance thermometers consist of a small crystal of doped germanium onto which four leads (two

current, two potential) are attached. These lead resistances are comparable with the sensor resistance and are similarly temperature dependent. This thermometer element is in a sealed jacket with a low pressure of exchange gas. Figure 7 shows the resistance–temperature characteristics for three of these resistors (labled GE), which are intended for different temperature ranges. The minimum usable temperature in each case is defined as that at which the resistance approaches $10^5\ \Omega$. The shapes of the calibration curves are quite similar, with, as a crude approximation, $d \ln R/d \ln T \simeq -2$. A detailed inspection of these relations reveals a complex behavior, with a nonmonotonic temperature dependence for dR/dT, so the generation of analytical expressions for the resistance–temperature characteristic is difficult.

Germanium resistance thermometers served as the basis for low-temperature standards thermometry for many years, until rhodium–iron thermometers were introduced. The major advantages of germanium resistance thermometers for experimental work are their relatively small size, high sensitivity, and good stability. While the higher-resistance thermometers can be used up to 77 K, they cannot be used at much higher temperatures because the temperature coefficient changes sign and is positive near room temperature. Their magnetoresistance is rather high and complex, and they are seldom used for measurements in large magnetic fields. For accurate work above roughly 30 K, dc and ac calibrations of these thermometers may differ significantly, dependent on the frequency, so the measurement method corresponding to the calibration must be used.

Thermistors are two-lead sintered metal–oxide devices of a generally small mass, much smaller than any of the above thermometers'. This, combined with the high sensitivity, is their major attraction. The extreme sensitivity requires that a thermistor be chosen to work in a specific temperature range, since otherwise the resistance will be either too small or too large. They have been used at temperatures from 4.2 K (seldom) to 700°C (special design). Their stability can be quite good, especially for the bead designs, when they are handled with care.

The forward voltage of semiconducting diodes also has a well-defined dependence on temperature, which has been used to produce thermometers that are small in size and dependable. Figure 8 gives the voltage–temperature relationships for silicon and gallium arsenide diode thermometers as obtained with a 10-μA measuring current. The gallium arsenide calibration is smoother than that for the silicon diode, with the knee in the silicon curve being rather sharp. At low temperatures, the sensitivity of these thermometers can be quite good (better than 1 mK), with an accuracy and reproducibility of 0.1 K or better. At higher temperatures, these limits should be increased by about an order of magnitude. Standard voltage–temperature relations for selected classes of these diodes allow interchange of off-the-shelf devices with anticipated low-temperature and high-temperature accuracies of 0.1 and 1 K, respectively.

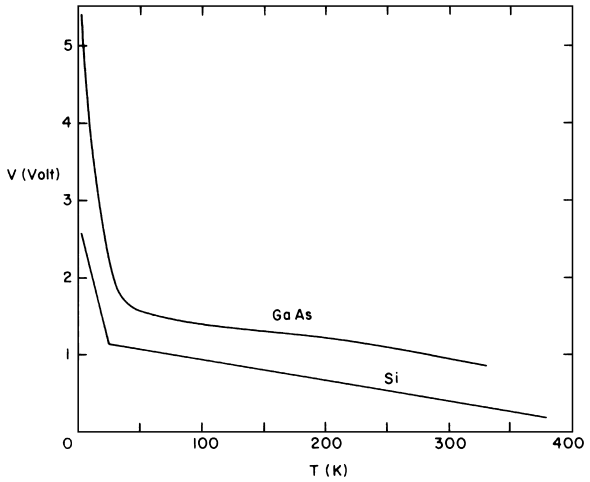


FIGURE 8 The temperature dependences of the forward voltages for two commercial diode thermometers. [Courtesy of Lake Shore Cryotronics, Inc.]

C. Thermocouples

The existence of a temperature gradient in a conductor will cause a corresponding emf to be generated in this conductor which depends on the gradient (the thermoelectric effect). While this emf (or voltage) cannot be measured directly for a single conductor, the difference between the thermal emfs for two materials can be measured and can be used to measure temperatures, as in a thermocouple. When two wires of dissimilar materials are joined at each end and the ends are kept at different temperatures, a (thermoelectric) voltage will appear across a break in the circuit. This voltage will depend on the temperature difference and, also, on the difference between the thermoelectric powers of the two materials. The temperature dependence of this voltage is called the "Seebeck coefficient."

The thermocouple which was used to define the high-temperature IPTS-68 interpolation relation (platinum–10% rhodium/platinum) gives the emf (E)-vs-temperature relation, labeled S in Fig. 9. Noble-metal thermocouples typically have a relatively low sensitivity (roughly 10 μV/K) and calibrations which may change with strain and annealing. These drawbacks are compensated by the usefulness of these thermocouples for work at very high-temperatures. In time, these traditional high-temperature thermocouples may be replaced by gold–platinum and/or platinum–palladium thermocouples, which have similar

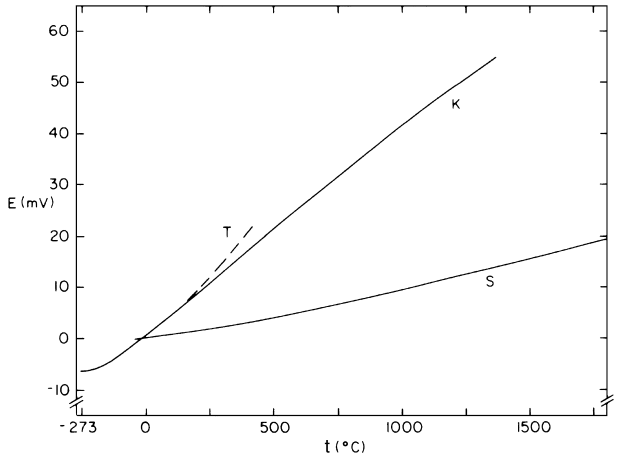


FIGURE 9 The voltage–temperature characteristics for typical noble-metal (S) and base-metal (K) thermocouples.

sensitivities but are more reproducible. More sensitive (basemetal) thermocouples are available for lower-temperature use, and two of these also are shown in Fig. 9. The type K (K) thermocouple uses nickel–chromium-vs-nickel–aluminum alloys, and the type T (T) uses copper vs a copper–nickel alloy. While Seebeck coefficients generally are very small below roughly 20 K, relatively large values (10 μV/K or so) are observed for dilute alloys (less than 0.1%) of iron in gold; these thermocouples are useful even below 1 K.

Thermocouples are convenient, especially when emfs are measured with modern semiconductor instrumentation. The reference junction generally is chosen to be at the ice point (0°C), where precautions must be taken if an ice bath is used. The junction must be electrically isolated from the bath to prevent leakage to ground, which could give false readings, and it must extend sufficiently far into the bath so that heat conduction along the wires to the junction is not important. Finally, the junction must be surrounded by melting ice (a mixture of ice and water), not cold water, since the density of water is minimum at 4°C and temperature gradients exist in water on which ice is floating. The ice bath can be replaced by an electronic device for which the output voltage simulates an ice bath and is independent of ambient temperature.

Thermocouples are relatively sensitive to their environment, and their calibration can be affected in many, sometimes subtle, ways. Annealing, oxidation, and alloying effects can change the Seebeck coefficient, while extraneous, emfs are introduced when strains and a temperature gradient coexist along a wire. Care clearly must be taken in experimental arrangements involving thermocouples, and the standard tables that exist for the various commonly used types of thermocouples must be applied judiciously. It is important to remember that the thermal produced by a thermocouple is developed along that part of the wire passing through a temperature gradient; it has nothing to do with the junction. Consequently, strains and inhomogeneities present in that part of the wire in the temperature gradient will lead to errors in the temperature measurement.

D. Optical Pyrometry

Some of the problems involved in optical pyrometry were addressed in an earlier section, with the emissivity of the source a major concern. Commercial pyrometers have been in use for many years and have been a part of the International Temperature Scales since 1927. Early optical pyrometers matched the brightness of the radiation source with that of a filament as the filament current was varied. The temperature of the source was then calibrated directly in terms of the current through the filament. Neutral density filters are used to extend the range of these pyrometers to higher temperatures. Considerable skill is required to use these "disappearing filament" pyrometers (the filament disappears in an image of the source) reproducibly, but they are used widely in industry.

The visual instruments have been replaced in standards and, also, in most practical applications by photoelectric pyrometers, in which a silicon diode detector or a photomultiplier tube replaces the eye as the detector. These instruments have a high sensitivity and can be used with interference filters to increase their accuracy [Eq. (3)]. A major concern in optical pyrometry is that real objects do not show ideal black-body radiation characteristics but have an emittance that differs from that of a black body in a manner that can be a function of the temperature, wavelength, and surface condition. Pyrometers that operate at two or more distinct wavelengths provide at least partial compensation for these effects.

A recent development in high-temperature optical pyrometry uses a fine sapphire fiber light pipe and photoelectric detection to obtain the temperature of a system that cannot be viewed directly. The end of the fiber may be encapsulated to form a black body (producing a self-contained thermometer) or the fiber may be used to view directly the object whose temperature is to be determined. Very sensitive semiconducting infrared detectors have made possible the use of total-radiation thermometers at and above room temperature for noncontact detection of temperature changes in processing operations and even, for instance, to determine the location of "heat leaks" in the insulation of a house. The slight excess temperature associated with certain tumors in medical applications has also been detected in this way.

E. Miscellaneous Thermometry

Many other thermometric systems are useful, some for specific applications. The variation with temperature of certain quartz piezoelectric coefficients gives a thermometer with a frequency readout. Very sensitive gas thermometers can be made with pressure changes sensed by changes in the resonant frequency of tunnel-diode circuits. Glass–ceramic capacitance thermometers are unique in that they have no magnetic field dependence, so are useful for low-temperature measurements in large magnetic fields.

Superconducting technology using SQUIDs allows the detection of very small changes in magnetic flux and, hence, in the current flowing through a loop of wire. Major advantages are the high sensitivity and the capability of using small samples in, for instance, magnetic thermometry and the measurement of low voltages. They have, for example, been used with gold–iron thermocouples for high-precision temperature measurements below 1 K. SQUIDs are primarily low-temperature devices but have been applied to routine measurements at room temperature and above.

Vapor pressure thermometry, with judicious choice of working substance, allows a very high sensitivity, but only, except at liquid helium temperatures, in a narrow temperature region. Here, capacitive diaphragm gauges and other modern pressure-sensing devices replace the conventional mercury manometer and allow remote readout of the pressures involved.

SEE ALSO THE FOLLOWING ARTICLES

CRITICAL DATA IN PHYSICS AND CHEMISTRY • CRYOGENICS • HEAT TRANSFER • THERMAL ANALYSIS • THERMODYNAMICS • THERMOELECTRICITY • TIME AND FREQUENCY

BIBLIOGRAPHY

American Institute of Physics (1992). "Temperature: Its Measurement and Control in Science and Industry," Vol. 6, Proceedings of the Symposium on Temperature, AIP, New York. (See also Vol. 5 in the same series.)

Bureau International des Poids et Mesures (BIPM) (1991). "Supplementary Information for the ITS-90," BIPM, Sevres, France. (A bibliography of recent articles on thermometry from national metrology institutes can be found at the BIPM web site: www.bipm.org.)

Bureau International des Poids et Mesures (BIPM) (1996). *Metrologia* 33, No. 4, 289–425 (a special issue devoted wholly to thermometry).

Hudson, R. P. (1980). "Measurement of temperature." *Rev. Sci. Instrum.* **51,** 871.

Quinn, T. J. (1990). "Temperature," 2nd ed., Academic Press, New York.



Underwater Acoustics

William A. Kuperman
University of California, San Diego

I. Ocean Acoustic Environment
II. Physical Mechanisms
III. Sonar Equation
IV. Sound Propagation Models
V. Quantitative Description of Propagation
VI. Sonar Array Processing
VII. Active Sonar Processing
VIII. Appendix: Units

GLOSSARY

Active sonar A sonar which emits sounds and receives its echo.
Beamforming Phasing an array to form a set of "look directions."
Convergence zone propagation Spatially periodic (≈35–65 km) refocusing of sound from a shallow source producing zones of high intensity near the surface due to the upward refracting nature of the sound speed profile and the absence of bottom interaction.
Decibels Ten times the logarithm in base 10 of a ratio of intensities.
Deep scattering layer A layer in the water column populated by organisms that scatter sound, and which typically undergoes diurnal variations in depth.
Deep sound channel A sound channel occurring in deep water whose axis is at the the minimum of the sound speed profile and in which propagation does not involve interaction with the ocean surface or bottom.
Matched field processing Beamforming by matching the data on an array with the solutions of the wave equation specific to the environment.
Passive sonar A sonar which only receives sound.
Propagation loss The ratio in decibels, between the acoustic intensity at a field point and the intensity at a reference distance (typically 1 m) from the source.
Reverberation The scattered acoustic field from an active sonar source which acts as interference in the sonar system.
Sound speed profile The speed of sound as a function of depth.
Surface duct A sound channel whose upper boundary is the ocean surface, formed when there is a local sound speed profile minimum near the ocean surface.
Transmission loss The negative of propagation loss.

IT IS WELL established that sound waves, rather than electromagnetic waves, propagate long distances in the ocean. Hence, in the ocean as opposed to air or a vacuum, there is SONAR (Sound Navigation and Ranging)

instead of radar, acoustic communication instead of radio, and acoustic imaging and tomography instead of microwave or optical imaging or X-ray tomography. Underwater acoustics is the science of sound in water (most commonly in the ocean) and encompasses not only the study of sound propagation, but also the masking of sound signal by interfering phenomena and the signal processing for extracting these signals from interference. This article will present the basic physics of ocean acoustics and then discuss applications. The decibel units used in underwater acoustics are described in the Appendix.

I. OCEAN ACOUSTIC ENVIRONMENT

The acoustic properties of the ocean such as the paths along which sound from a localized source travel are mainly dependent on the ocean sound speed structure, which in turn is dependent on the oceanographic environment. The combination of water column and bottom properties leads to a set of generic sound propagation paths descriptive of most propagation phenomena in the ocean.

A. Ocean Environment

Sound speed in the ocean water column is a function of temperature, salinity, and ambient pressure. Since the ambient pressure is a function of depth, it is customary to express the sound speed (c) in meters per second as an empirical function of temperature (T) in degrees centigrade, salinity (S) in parts per thousand, and depth (z) in meters, for example,

$$c = 1449.2 + 4.6T - 0.055T^2 + 0.00029T^3 + (1.34 - 0.01T)(S - 35) + 0.016z. \quad (1)$$

Figure 1 shows a typical set of sound speed profiles, indicating greatest variability near the surface. In a warmer season (or warmer part of the day, sometimes referred to as the "afternoon effect"), the temperature increases near the surface and hence the sound speed increases toward the sea surface. In nonpolar regions where mixing near the surface due to wind and wave activity is important, a *mixed layer* of almost constant temperature is often created. In this isothermal layer sound speed increases with depth because of the increasing ambient pressure, the last term in Eq. (1). This is the *surface duct* region. Below the mixed layer is the thermocline where the temperature and hence the sound speed decreases with depth. Below the thermocline, the temperature is constant and the sound speed increases because of increasing ambient pressure. Therefore, between the deep isothermal region and the mixed

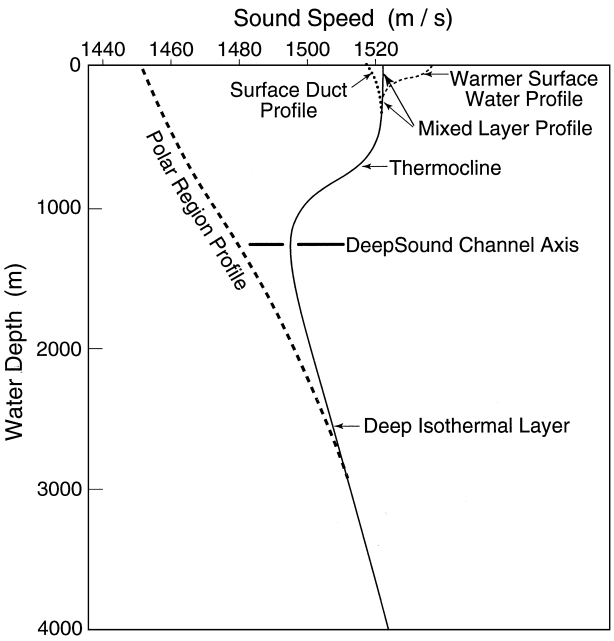


FIGURE 1 Generic sound speed profiles.

layer, there is a depth at minimum sound speed referred to as the axis of the *deep sound channel*. However, in polar regions, the water is coldest near the surface, so that the minimum sound speed is at the surface. Figure 2 is a contour display of the sound speed structure of the North and South Atlantic with the deep sound channel axis indicated by the heavy dashed line. Note the deep sound channel becomes shallower toward the poles. Aside from sound speed effects, the ocean volume is absorbtive and will cause attenuation that increases with acoustic frequency.

Shallower water such as that in continental shelf and slope regions is not deep enough for the depth-pressure term in Eq. (1) to be significant. Thus the winter profile tends to isovelocity simply because of mixing, whereas the summer profile has a higher sound speed near the surface due to heating; both are schematically represented in Fig. 3.

The sound speed structure regulates the interaction of sound with the boundaries. The ocean is bounded above by air which is a perfect reflector; however, it is often rough, causing sound to scatter in directions away from the "specular" reflecting angle. The ocean bottom is typically a complicated, rough, layered structure supporting elastic waves. Its geoacoustic properties are summarized by density, compressional and shear speed, and attenuation profiles. The two basic interfaces, air/sea and sea/bottom, can be thought of as the boundaries of an acoustic waveguide whose internal index of refraction is determined by the fundamental oceanographic parameters represented in the sound speed equation, Eq. (1).

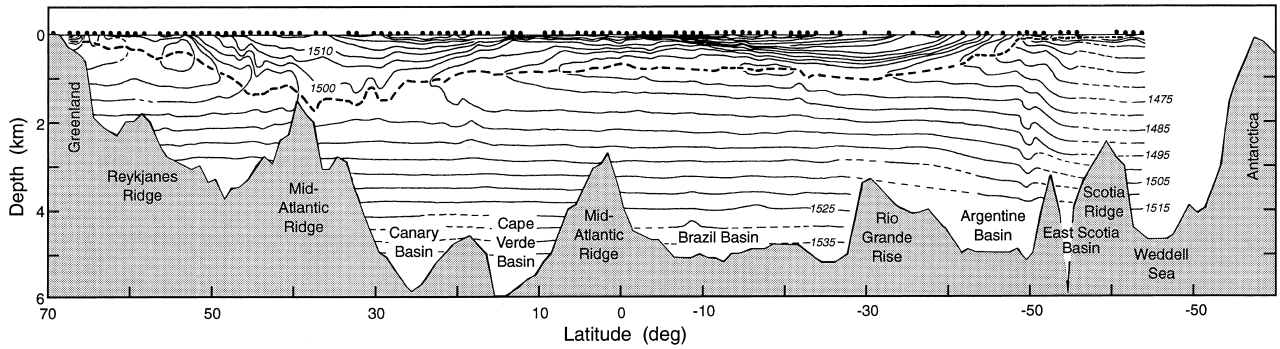


FIGURE 2 Sound speed contours at 5 m/sec intervals taken from the North and South Atlantic along 30.50° *W*. Dashed line indicates axis of deep sound channel (from Northrup 1974).

B. Basic Acoustic Propagation Paths

Sound propagation in the ocean can be qualititatively broken down into three classes: very short range, deep water, and shallow water propagation.

1. Very Short Range Propagation

The amplitude of a point source in free space falls off with range r as r^{-1}; this geometric loss is called *spherical spreading*. Most sources of interest in the deep ocean are nearer the surface than the bottom. Hence, the two main short range paths are the direct path and the surface reflected path. When these two paths *interfere*, they produce a spatial distribution of sound oftened referred to as a Lloyd mirror pattern, as shown in the inset of Fig. 4. Also, with reference to Fig. 4, note that *transmission loss* is a decibel measure of relative intensity (see Appendix), the latter being proportional to the square of the acoustic amplitude.

2. Long Range Propagation Paths

Figure 5 is a schematic of propagation paths in the ocean resulting from the sound speed profiles (indicated by the dashed line) described above in Fig. 1. These paths can be understood from *Snell's law*,

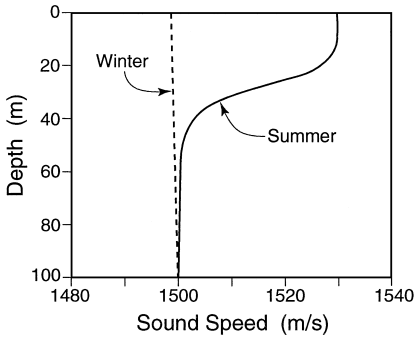


FIGURE 3 Typical summer and winter shallow water sound speed profiles.

$$\frac{\cos\theta(z)}{c(z)} = constant, \tag{2}$$

which relates the ray angle $\theta(z)$, with respect to the horizontal, to the local sound speed $c(z)$ at depth z. The equation requires that the higher the sound speed, the smaller the angle with the horizontal, meaning, that sound bends away from regions of high sound speed; or said another way, sound bends toward regions of low sound speed. Therefore, paths 1, 2, and 3 are the simplest to explain since they are paths that oscillate about the local sound speed minima. For example, path 3, depicted by a ray leaving a source near the deep sound channel axis at a small horizontal angle, propagates in the deep sound channel. This path, in temperate lattitudes where the sound speed minimum is far from the surface, permits propagation over distances of thousands of kilometers. Path 4, which is at slightly steeper angles and is usually excited by a near surface source, is *convergence zone* propagation, a spatially periodic (35–65 km) refocusing phenomenon producing zones of high intensity near the surface due to the upward refracting nature of the deep sound-speed profile. Regions in between these zones are referred to as shadow regions. Referring back to Fig. 1, there may be a depth in the deep isothermal layer at which the sound speed is the same at it is at the surface; this depth is called the *critical depth* and is the lower limit of the deep sound channel. A positive critical depth specifies that the environment supports long distance propagation without bottom interaction, whereas a negative critical depth specifies that the ocean bottom is the lower boundary of the deep sound channel. The *bottom bounce* path 5 is also a periodic phenomenon but with a shorter cycle distance and shorter propagation distance because of losses when sound is reflected from the ocean bottom.

3. Shallow Water and Waveguide Propagation

In general, the ocean can be thought of as an acoustic waveguide; this waveguide physics is particularly evident

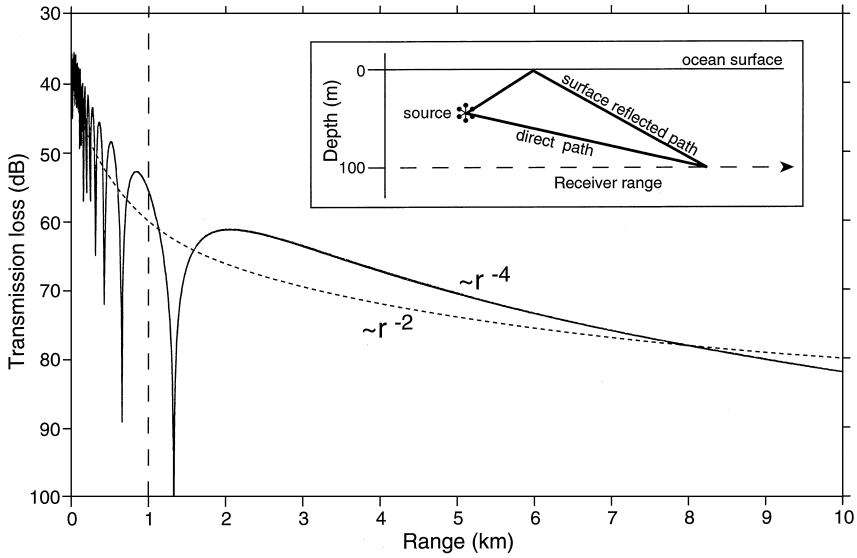


FIGURE 4 The inset shows the geometry of the Lloyd mirror effect. The plots show a comparison of Lloyd mirror to spherical spreading. Transmission losses are plotted in decibels corresponding to losses of $10 \log r^2$ and $10 \log r^4$, respectively, as explained in Section I.C.

in shallow water (inshore out to the continental slope, typically to depths of a few hundred meters). Snell's law applied to the summer profile in Fig. 3 produces rays which bend more toward the bottom than winter profiles in which the rays tend to be straight. This implies two effects with respect to the ocean bottom: (1) For a given range, there are more bounces off the ocean bottom in the summer than in the winter; (2) the ray angles intercepting the bottom are steeper in the summer than in the winter. A qualitative understanding of the reflection properties of the ocean bottom should therefore be very revealing of sound propagation in summer versus winter. Basically, near-grazing incidence is much less lossy than larger, more vertical angles of incidence. Since summer propagation paths have more bounces, each of which is at steeper angles than those of winter paths, summer shallow water propagation is lossier than in winter. This result is tempered by rough winter surface conditions that generate large scattering losses at the higher frequencies.

For simplicity, we consider an isovelocity waveguide bounded above by the air/water interface and below by a two-fluid interface. From Section II.C., we have perfect reflection with a 180-degree phase change at the surface, and for paths more horizontal than the bottom critical angle, there will also be perfect bottom reflection. Therefore, as schematically indicated in Fig. 6a, ray paths within a cone of $2\theta_c$ will propagate unattenuated down the waveguide. Because the upgoing and downgoing rays have equal amplitudes, preferred angles will exist such that perfect constructive interference can occur. These particular angles can be associated with the normal modes of

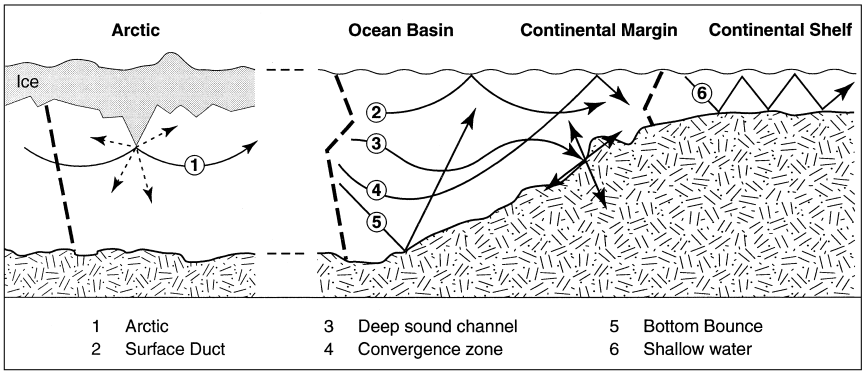


FIGURE 5 Schematic representation of various types of sound propagation in the ocean.

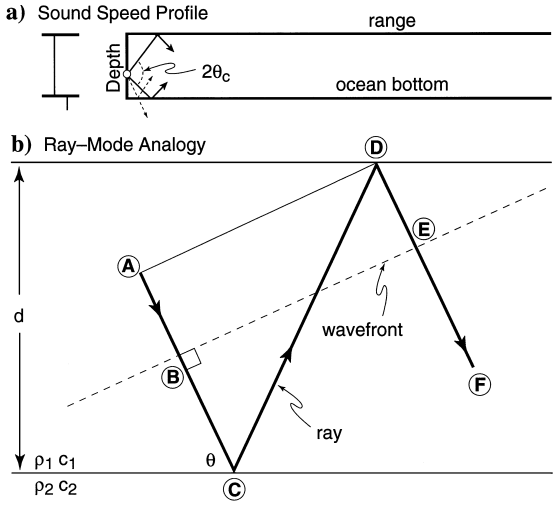


FIGURE 6 Ocean waveguide propagation. (a) Long distance propagation occurs within a cone of $2\theta_c$. (b) There are a discrete set of paths that reflect off the bottom and surface that constructively interfere. For the example shown, the condition for constructive interference is that the phase change along BCDE be a multiple of two π.

the waveguide as formally derived from the wave equation in Section IV. However, it is instructive to understand the geometric origin of the waveguide modal structure. Figure 6b is a schematic of a ray reflected from the bottom and then the surface of a "Pekeris" waveguide (an environment with constant sound speeds and densities in the water column and fluid bottom, respectively). Consider a ray along the path ACDF and its wavefront which is perpendicular to the ray. The two downgoing rays of equal amplitude, AC and DF, will constructively interfere if points B and E have a phase difference of an integral number of 360 degrees (and similarly for upgoing rays). There will be a discrete set of angles up to the critical angle for which this constructive interference takes place and, hence, for which sound propagates. This discrete set, in terms of wave physics, is called the normal modes of the waveguide and is further discussed in Section IV.D.

C. Geometric Spreading Loss

The energy per unit time emitted by a sound source is flowing through a larger area with increasing range. Intensity is the the power flux through a unit area which translates to the energy flow per unit time through a unit area. The simplest example of geometric loss is spherical spreading for a point source in free space where the area increases as $4\pi r^2$, where r is the range from the point source. So spherical spreading results in an intensity decay proportional to r^{-2}. Since intensity is proportional to the square of the pressure amplitude, the fluctuations in pressure induced by the sound, p, decay as r^{-1}. For range independent ducted propagation, that is, where rays are refracted or reflected back toward the horizontal direction, there is no loss associated with the vertical dimension. In this case, the spreading surface is the area of cylinder whose axis is in the vertical direction passing through the source, $2\pi r H$, where H is the depth of the duct (waveguide) and is constant. Geometric loss in the near field Lloyd mirror regime requires consideration of interfering beams from direct and surface reflected paths. To summarize, the geometric spreading laws for the pressure field (recall that intensity is proportional to the sqaure of the pressure) are:

- Spherical spreading loss: $p \propto r^{-1}$
- Cylindrical spreading loss: $p \propto r^{-1/2}$
- Lloyd mirror loss: $p \propto r^{-2}$.

II. PHYSICAL MECHANISMS

The physical mechanisms associated with the generation, reception, attenuation, and scattering of sound in the ocean are discussed in this section.

A. Transducers

A transducer converts some sort of energy to sound (source) or converts sound energy (receiver) to an electrical signal. In underwater acoustics, piezoelectric and magnetostrictive transducers are commonly used; the former connects electric polarization to mechanical strain and the latter connects magnetization of a ferromagnetic material to mechanical strain. In addition there are: electrodynamic transducers in which sound pressure oscillations move a current-carrying coil through a magnetic field causing a back electromagnetic field, and electrostatic transducers in which charged electrodes moving in a sound field change the capacitance of the system. Explosion, airgun, electric discharge, and lasers are also used as wideband sources.

B. Volume Attenuation

Volume attenuation increases with frequency. In Fig. 5, the losses associated with path 3 only include volume attenuation and scattering, because this path does not involve boundary interactions. The volume scattering can be biological in origin or arise from interaction with internal wave activity in the vicinity of the upper part of the deep sound channel where paths are refracted before they would interact with the surface. Both of these effects are small at low frequencies. This same internal wave region is also on the lower boundary of the surface duct, allowing scattering out of the surface duct, thereby also constituting a loss mechanism for the surface duct. This mechanism also

leaks sound into the deep sound channel, a region which without scattering would be a shadow zone for a surface duct source. This type of scattering from internal waves is also a source of fluctuation of the sound field.

Attenuation is characterized by an exponential decay of the sound field. If A_0 is the rms amplitude of the sound field at unit distance from the source, then the attenuation of the sound field causes the amplitude to decay with distance along the path, r:

$$A = A_0 \exp(-\alpha r), \tag{3}$$

where the unit of α is nepers/distance. The attenuation coefficient can be expressed in decibels per unit distance by the conversion $\alpha' = 8.686\alpha$. The frequency dependence of attenuation can be roughly divided into four regimes as displayed in Fig. 7. In Region I, leakage out of the sound channel is believed to be the main cause of attenuation. The main mechanisms associated with Regions II and III are boric acid and magnesium sulfate chemical relaxation. Region IV is dominated by the shear and bulk viscosity associated with fresh water. A summary of the approximate frequency dependence (f in kHz) of attenuation (in units of dB/km) is given by

$$\alpha'(dB/km) = 3.3 \times 10^{-3} + \frac{0.11 f^2}{1 + f^2} + \frac{43 f^2}{4100 + f^2} + 2.98 \times 10^{-4} f^2, \tag{4}$$

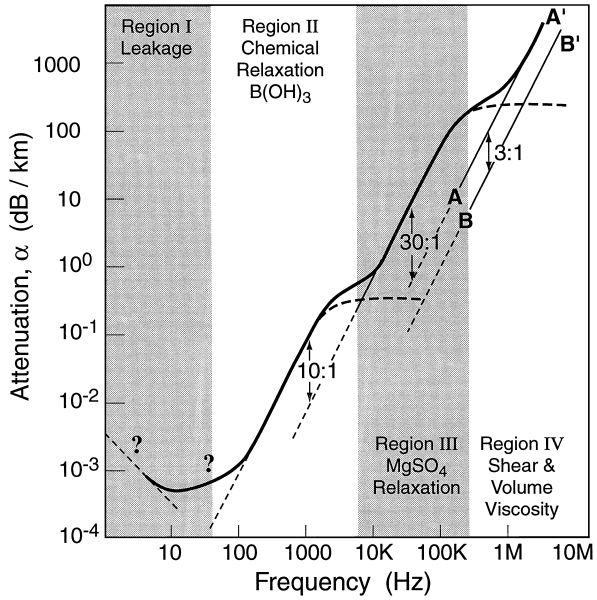


FIGURE 7 Regions of different dominent processes at attenuation of sound in seawater [From Urick, R. J. (1979). Sound Propagation in the Sea. Washington: U.S. G.P.O.]. The attenuation is given in dB per kiloyard.

with the terms sequentially associated with Regions I–IV in Fig. 7.

C. Bottom Loss

The structure of the ocean bottom affects those acoustic paths which interact with the ocean bottom. This bottom interaction is summarized by *bottom reflectivity*, the amplitude ratio of reflected and incident plane waves at the ocean-bottom interface as a function of grazing angle, θ (see Fig. 8a). For a simple bottom which can be represented by a semi-infinite half-space with constant sound speed c_b and density ρ_b, the reflectivity is given by

$$\mathcal{R}(\theta) = \frac{\rho_b k_{wz} - \rho_w k_{bz}}{\rho_b k_{wz} + \rho_w k_{bz}}, \tag{5}$$

with the subscript w denoting water; the wavenumbers are given by

$$k_{iz} = (\omega/c_i)\sin\theta_i \equiv k \sin\theta_i; \quad i = w, b. \tag{6}$$

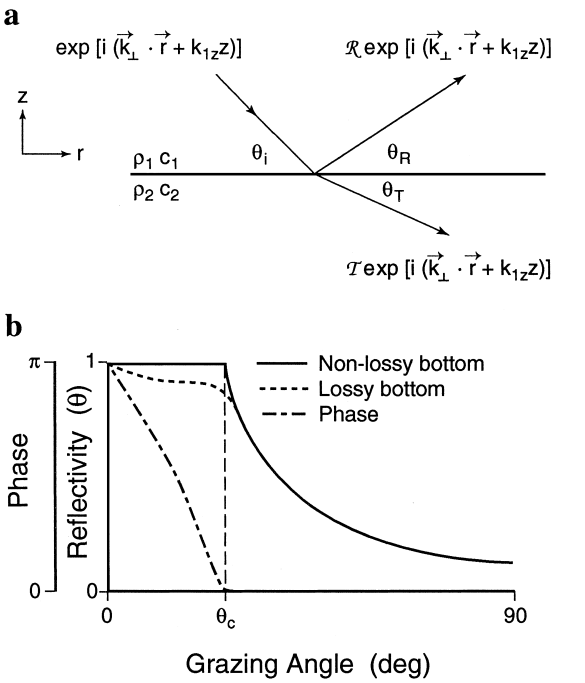


FIGURE 8 The reflection and transmission process. Grazing angles are defined relative to the horizontal. (a) A plane wave is incident on an interface separating two media with densities and sound speeds ρ, c. $\mathcal{R}(\theta)$ and $\mathcal{T}(\theta)$ are reflection and transmission coefficients. Snell's law is a statement that $k_\perp$, the horizontal component of the wave vector, is the same for all three waves. (b) Rayleigh reflection curve (Eq. 5) as a function of the grazing angle (θ in (a)) indicating critical angle θ_c. The dashed curve shows that if the second medium is lossy, there is less than perfect reflection below the critical angle. Note that for the nonlossy, bottom there is complete reflection below the critical angle, but with a phase change.

The incident and transmitted grazing angles are related by Snell's law,

$$c_b \cos\theta_w = c_w \cos\theta_b, \tag{7}$$

and the incident grazing angle θ_w is also equal to the angle of the reflected plane wave.

For this simple water-bottom interface for which we take $c_b > c_w$, there exists a critical grazing angle θ_c below which there is perfect reflection,

$$\cos\theta_c = \frac{c_w}{c_b}. \tag{8}$$

For a lossy bottom, there is no perfect reflection, as also indicated in a typical reflection curve in Fig. 8b. These results are approximately frequency independent. However, for a layered bottom, the reflectivity has a complicated frequency dependence. It should be pointed out that if the density of the second medium vanishes, the reflectivity reduces to the pressure release case of $\mathcal{R}(\theta) = -1$.

D. Scattering and Reverberation

Scattering caused by rough boundaries or volume inhomogeneities is a mechanism for loss (attenuation), reverberant interference, and fluctuation. Attenuation from volume scattering is addressed in Section II.C. In most cases, it is the mean or coherent (or specular) part of the acoustic field which is of interest for a sonar or communications application, and scattering causes part of the acoustic field to be randomized. Rough surface scattering out of the "specular direction" can be thought of as an attenuation of the mean acoustic field, and typically increases with increasing frequency. A formula often used to describe reflectivity from a rough boundary is

$$\mathcal{R}'(\theta) = \mathcal{R}(\theta)\exp\left(-\frac{\Gamma^2}{2}\right), \tag{9}$$

where $\mathcal{R}(\theta)$ is the reflection coefficient of the smooth interface and Γ is the Rayleigh roughness parameter defined as $\Gamma \equiv 2k\sigma \sin\theta$ where $k = 2\pi/\lambda$, λ is the acoustic wavelength, and σ is the rms roughness (height).

The scattered field is often referred to as reverberation. Surface, bottom or volume scattering strength, $S_{S,B,V}$, is a simple parameterization of the production of reverberation and is defined as the ratio in decibels of the sound scattered by a unit surface area or volume referenced to a unit distance, I_{scat}, to the incident plane wave intensity, I_{inc},

$$S_{S,B,V} = 10\log\frac{I_{scat}}{I_{inc}}. \tag{10}$$

The Chapman–Harris curves predicts the ocean surface scattering strength in the 400–6400 Hz region,

$$S_S = 3.3\beta\log\frac{\theta}{30} - 42.4\log\beta + 2.6;$$
$$\beta = 107(wf^{1/3})^{-0.58}, \tag{11}$$

where θ is the grazing angle in degrees, w the wind speed in m/sec, and f the frequency in Hz.

The simple characterization of bottom backscattering strength utilizes Lambert's rule for diffuse scattering,

$$S_B = A + 10\log\sin^2\theta \tag{12}$$

where the first term is determined empirically. Under the assumbtion that all incident energy is scattered into the water column with no transmission into the bottom, A is -5 dB. Typical realistic values for A which have been measured are -17 dB for big Basalt Mid-Atlantic Ridge cliffs and -27 dB for sediment ponds.

Volume scattering strength is typically reduced to a surface scattering strength by taking S_V as an average volume scattering strength within some layer at a particular depth; then the corresponding surface scattering strength is

$$S_S = S_V + 10\log H \tag{13}$$

where H is the layer thickness. The column or integrated scattering strength is defined as the case for which H is the total water depth.

Volume scattering usally decreases with depth (about 5 dB per 300 m) with the exception of the deep scattering layer. For frequencies less than 10 kHz, fish with air-filled swim bladders are the main scatterers. Above 20 kHz, zooplankton or smaller animals that feed upon phytoplankton and the associated biological chain are the scatterers. The deep scattering layer (DSL) is deeper in the day than in the night, changing most rapidly during sunset and sunrise. This layer produces a strong scattering increase of 5–15 dB within 100 m of the surface at night and virtually no scattering in the daytime at the surface since it migrates down to hundreds of meters. Since higher pressure compresses the fish swim bladder, the backscattering acoustic resonance tends to be at a higher frequency during the day when the DSL migrates to greater depths. Examples of day and night scattering strengths are shown in Fig. 9.

Finally, near-surface bubbles and bubble clouds can be thought of as either volume or surface scattering mechanisms acting in concert with the rough surface. Bubbles have resonances (typically greater than 10 kHz) and at these resonances, scattering is strongly enhanced. Bubble clouds have collective properties; among these properties is that a bubbly mixture, as specified by its void fraction (total bubble gas volume divided by water volume), has a considerably lower sound speed than water.

E. Ambient Noise

There are essentially two types of ocean acoustic noise: manmade and natural. Generally, shipping is the most important source of manmade noise, though noise from offshore oil rigs is becoming more and more prevalent.

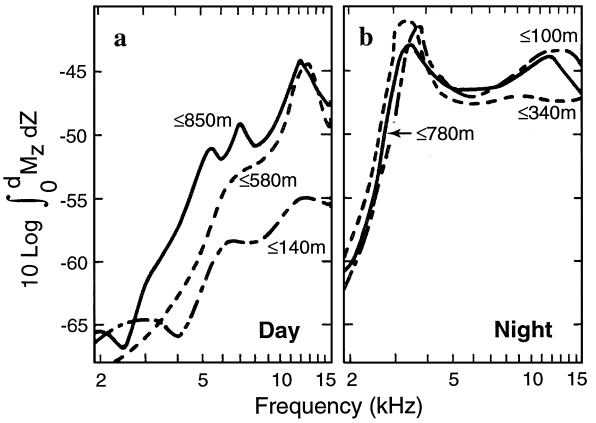


FIGURE 9 Day and night scattering strength measurements using an explosive source as a function of frequency [from Chapman and Marshall (1966)]. The spectra measured at various times after the explosion are labeled with the depth of the nearest scatterer that could have contributed to the reverberation. The ordinate corresponds to S_V in Eq. (13). [From Chapman, R. P. and Harris, H. H. (1962). "Surface backscattering strengths measured with explosive sound sources," *J. Acoust. Soc. Am.* **34,** 1592–1597.]

Typically, natural noise dominates at low frequencies (below 10 Hz) and high frequencies (above a few hundred Hz). Shipping fills in the region between 10 and a few hundred Hz. A summary of the spectrum of noise is shown in Fig. 10. The higher frequency noise is usually parameterized according to sea state (also Beaufort number) and/or wind. Table I summarizes the description of sea state.

The sound speed profile affects the vertical and angular distribution of noise in the deep ocean. When there is a positive critical depth (see Section I.B.), sound from surface sources can travel long distances without interacting with the ocean bottom, but a receiver below this critical depth should sense less surface noise because propagation involves interaction with lossy boundaries, surface and/or bottom. This is illustrated in Fig. 11, which shows a deep water environment with measured ambient noise. Figure 12 is an example of vertical directivity of noise which also follows the propagation physics discussed above. The shallower depth is at the axis of the deep sound channel while the other is at the critical depth. The pattern is narrower at the critical depth where the sound paths tend to be horizontal since the rays are turning around at the lower boundary of the deep sound channel.

In a range independent ocean, Snell's law predicts a horizontal noise notch at depths where the speed of sound is less than the near-surface sound speed. Returning to Eq. (2), and reading off the sound speeds from Fig. 11 at the surface ($c = 1530$ m/sec) and say, 300 m (1500 m/sec), a horizontal ray ($\theta = 0$) launched from ocean surface would have an angle with respect to the horizontal of about 11°

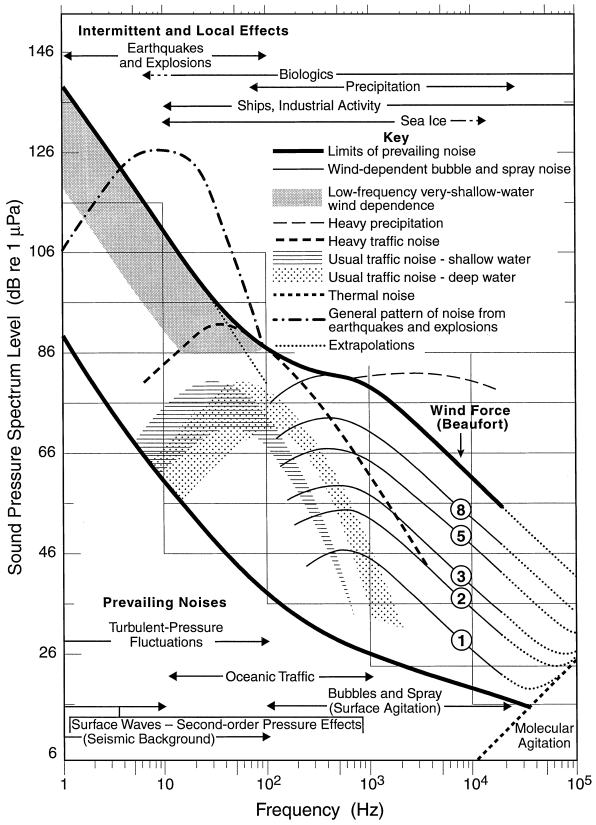


FIGURE 10 Composite of ambient noise spectra [From Wenz, G. M. (1962). "Acoustic ambient noise in the ocean: Spectra and sources," *J. Acoust. Soc. Am.* **34,** 1936–1956].

at 300 m depth. All other rays would arrive with greater vertical angles. Hence we expect this horizontal notch. However, the horizontal notch is often not seen at shipping noise frequencies. That is because shipping tends to be concentrated in continental shelf regions, and propagation down a continental slope converts high angles rays to lower angles at each bounce. There are also deep sound channel shoaling effects that result in the same trend in angle conversion.

III. SONAR EQUATION

A major application of underwater acoustics is sonar system technology. The performance of a sonar is often approximately described simply in terms of the sonar equation. The methodology of the sonar equation is analogous to an accounting procedure involving acoustic signal, interference, and system characteristics.

A. Passive Sonar Equation

A passive sonar system uses the radiated sound from a target to detect and locate the target. A radiating object

TABLE I Descriptions of the Ocean Sea Surface

Sea criteria	Wind speed			12-hr wind	Fully arisen sea			
	Beaufort scale	Range knots (m/s)	Mean knots (m/s)	Wave height[a,b] ft (m)	Wave height[a,b] ft (m)	Duration[b,c] hr	Fetch[b,c] naut. miles (km)	Seastate scale
Mirrorlike	0	<1 (<0.5)						0
Ripples	1	1–3 (0.5–1.7)	2 (1.1)					1/2
Small wavelets	2	4–6 (1.8–3.3)	5 (2.5)	<1 (<0.30)	<1 (<0.30)			1
Large wavelets, scattered whitecaps	3	7–10 (3.4–5.4)	8-1/2 (4.4)	1–2 (0.30–0.61)	1–2 (0.30–0.61)	<2.5	<10 (<19)	2
Small waves, frequent whitecaps	4	11–16 (5.5–8.4)	13-1/2 (6.9)	2–5 (0.61–1.5)	2–6 (0.61–1.8)	2.5–6.5	10–40 (19–74)	3
Moderate waves, many whitecaps	5	17–21 (8.5–11.1)	19 (9.8)	5–8 (1.5–2.4)	6–10 (1.8–3.0)	6.5–11	40–100 (74–185)	4
Large waves, whitecaps everywhere, spray	6	22–27 (11.2–14.1)	24-1/2 (12.6)	8–12 (2.4–3.7)	10–17 (3.0–5.2)	11–18	100–200 (185–370)	5
Heaped-up sea, blown spray, streaks	7	28–33 (14.2–17.2)	30-1/2 (15.7)	12–17 (3.7–5.2)	17–26 (5.2–7.9)	18–29	200–400 (370–740)	6
Moderately high, long waves, spindrift	8	34–40 (17.3–20.8)	37 19.0)	17–24 (5.2–7.3)	26–39 (7.9–11.9)	29–42	400–700 (740–1300)	7

[a] The average height of the highest one-third of the waves (significant wave height).
[b] Estimated from data given in U.S. Hydrographic Office (Washington, DC) publications HO 604 (1951) and HO 603 1955).
[c] The minimum fetch and duration of the wind needed to generate a fully arisen sea.
Note. Approximate relation between scales of wind speed, wave height, and sea state [From Wenz, G. M. (1962). "Acoustic ambient noise in the ocean: Spectra and sources," *J. Acoust. Soc. Am.* **34,** 1936–1956].

of source level *SL* (all units are in decibels) is received at a hydrophone of a sonar system at a lower signal level *S* because of the transmission loss "TL" it suffers (e.g., cylindrical spreading plus attenuation or a TL computed from one of the propagation models of Section IV),

$$S = SL - TL. \tag{14}$$

The noise, N, at a single hydrophone is subtracted from Eq. (14) to obtain the signal-to-noise ratio at a single hydrophone,

$$SNR = SL - TL - N. \tag{15}$$

Typically, a sonar system consists of an array or antenna of hydrophones which provides signal-to-noise enhancement through a beamforming process (see Section VI). This process is quantified in decibels by array gain *AG* (see Section VI.B.) that is added to the single hydrophone SNR to give the SNR at the output of the beamformer,

$$SNR_{BF} = SL - TL - N + AG. \tag{16}$$

Because detection involves addtional factors including sonar operator ability, it is necessary to specify a detection threshold, *DT* level above the SNR_{BF} at which there is a 50% (by convention) probability of detection. The difference between these two quantities is called signal excess (SE),

$$SE = SL - TL - N + AG - DT. \tag{17}$$

This decibel bookkeeping leads to an important sonar engineering descriptor called the figure of merit, FOM, which is the transmission loss that gives a zero signal excess,

$$FOM = SL - N + AG - DT \tag{18}$$

The FOM encompasses the various parameters a sonar engineer must deal with: expected source level, the noise environment, array gain, and the detection threshold. Conversely, since the FOM is a transmission loss, one can use the output of a propagation model (or if appropriate, a simple geometric loss plus attenuation) to estimate the minimum range at which a 50% probability of detection can be expected. This range changes with oceanographic conditions and is often referred to as the "range of the day" in navy sonar applications.

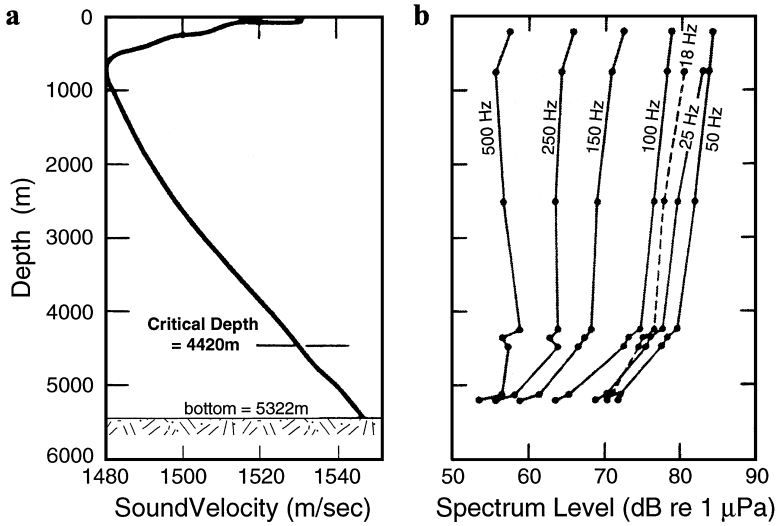


FIGURE 11 Noise in the deep ocean. (a) Sound speed profile and (b) noise level as a function of depth in the Pacific [From Morris, G. B. (1978). "Depth dependence of ambient noise in the Northeastern Pacific Ocean," *J. Acoust. Soc. Am.* **64,** 581–590].

B. Active Sonar Equation

A monostatic active sonar transmits a pulse to a target and its echo is detected at a receiver colocated with the transmitter. A bistatic active sonar has the receiver in a different location than the transmitter. The main differences between the passive and active cases is the addition of a target strength term, *TS*; reverberation and hence reverberation level, RL, is usually the dominant source of interference as opposed noise; and the transmission loss is over two paths: transmitter to target and target to receiver. In the monostatic case, the transmission loss is *2TL* where *TL* is the-one way transmission loss, and in the bistatic case, the transmission loss is the sum (in dB) over paths from the transmitter to the target and the target to the reciever, $TL_1 + TL_2$. The concept of the detection threshold is useful for both passive and active sonars. Hence, for signal excess, we have

$$SE = SL - TL_1 + TS - TL_2 - (RL + N) + AG - DT. \quad (19)$$

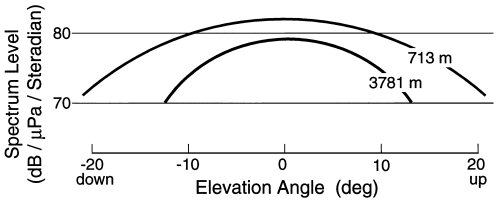


FIGURE 12 The vertical directionality of noise at the axis of the deep sound channel and at the critical depth in the Pacific [From Anderson, V. C. (1979). "Variations of the vertical directivity of noise with depth in the North Pacific," *J. Acoust. Soc. Am.* **66,** 1446–1452].

The corresponding FOM for an active system is defined for the maximum allowable two-way transmission loss with $TS = 0$ dB.

IV. SOUND PROPAGATION MODELS

The wave equation describing sound propagation is derived from the equations of hydrodynamics and its coefficients, and boundary conditions are descriptive of the ocean environment. There are essentially four types of models (computer solutions to the wave equation) to describe sound propagation in the sea:

1. Ray theory
2. The spectral method or fast field program (FFP)
3. Normal mode (NM)
4. Parabolic equation (PE).

All of these models allow for the fact that the ocean environment varies with depth. A model that also takes into account horizontal variations in the environment (i.e., sloping bottom or spatially variable oceanography) is termed range dependent. For high frequencies (a few kilohertz or above), ray theory is the most practical. The other three model types are more applicable and usable at lower frequencies (below a kilohertz). The models discussed here are essentially two-dimensional models, since the index of refraction has much stronger dependence on depth than on horizontal distance. Nevertheless, bottom topography and strong ocean features can cause horizontal refraction (out of the range-depth plane). Ray models are most easily

extendable to include this added complexity. Full three-dimensional wave models are extremely computationally intensive. A compromise that often works for "weak" three-dimensional problems is the "$N \times 2D$" approximation that combines two-dimensional solutions along radials to produce a three-dimensional solution.

A. The Wave Equation and Boundary Conditions

The wave equation for pressure, p, in cylindrical coordinates with the range coordinates denoted by $\mathbf{r} = (x, y)$ and the depth coordinate denoted by z (taken positive downward) for a source free region is

$$\nabla^2 p(\mathbf{r}, z, t) - \frac{1}{c^2(\mathbf{r}, z)} \frac{\partial^2 p(\mathbf{r}, z, t)}{\partial t^2} = 0, \qquad (20)$$

where $c(\mathbf{r}, z)$ is the sound speed in the wave propagating medium.

It is convenient to solve Eq. (20) in the frequency domain by assuming a solution with a frequency dependence of $\exp(-i\omega t)$ to obtain the Helmholtz equation ($K \equiv \omega/c$),

$$\nabla^2 p(\mathbf{r}, z) + K^2 p(\mathbf{r}, z) = 0, \qquad (21)$$

with

$$K^2(\mathbf{r}, z) = \frac{\omega^2}{c^2(\mathbf{r}, z)}. \qquad (22)$$

The range-dependent environment manifests itself as the coefficient $K^2(\mathbf{r}, z)$ of the partial differential equation for the appropriate sound speed profile. The range-dependent bottom type and topography appears as boundary conditions. In underwater acoustics, both fluid and elastic (shear supporting sediments and bottom strata) media are of interest. For simplicity we only consider fluid media below.

The most common plane interface boundary conditions encountered in underwater acoustics are the pressure release condition at the ocean surface,

$$p = 0, \qquad (23)$$

and at the ocean-bottom interface, the continuity of pressure

$$p_1 = p_2, \qquad (24)$$

and vertical particle velocity

$$\frac{1}{\rho_1} \frac{\partial p_1}{\partial z} = \frac{1}{\rho_2} \frac{\partial p_2}{\partial z}, \qquad (25)$$

where the ρ_i's are the densities of the two media. These latter boundary conditions applied to the plane wave fields in Fig. 8a yield the Rayleigh reflection coefficient given by Eq. (1).

The Helmholtz equation for an acoustic field from a point source is

$$\nabla^2 G(\mathbf{r}, z) + K^2(\mathbf{r}, z) G(\mathbf{r}, z) = -\delta^2(\mathbf{r} - \mathbf{r}_s)\delta(z - z_s), \qquad (26)$$

where the subscript "s" denotes the source coordinates. The acoustic field from a point source, $G(\mathbf{r})$, is obtained either by solving the boundary value problem of Eq. (26) (spectral method or normal modes) or by approximating Eq. (21) by an initial value problem (ray theory, parabolic equation).

B. Ray Theory

Ray theory is a geometrical, high-frequency approximate solution to Eq. (21) of the form

$$G(\mathbf{R}) = A(\mathbf{R}) \exp[i S(\mathbf{R})], \qquad (27)$$

where the exponential term allows for rapid variations as a function of range and $A(\mathbf{R})$ is a more slowly varying "envelope" which incorporates both geometrical spreading and loss mechanisms. The geometrical approximation is that the amplitude varies slowly with range (i.e., $(1/A)\nabla^2 A \ll K^2$) so that Eq. (21) yields the eikonal equation

$$(\nabla S)^2 = K^2. \qquad (28)$$

The ray trajectories are perpendicular to surfaces of constant phase (wavefronts), S, and may be expressed mathematically as follows:

$$\frac{d}{dl}\left[K \frac{d\mathbf{R}}{dl}\right] = \nabla K, \qquad (29)$$

where l is the arc length along the direction of the ray, and $\mathbf{R}$ is the displacement vector. The direction of average flux (energy) follows that of the trajectories, and the amplitude of the field at any point can be obtained from the density of rays.

The ray theory method is computationally rapid and extends to range-dependent problems. Furthermore, the ray traces give a physical picture of the acoustic paths. It is helpful in describing how sound redistributes itself when propagating long distances over paths that include shallow and deep environments and/or mid-latitude to polar regions. The disadvantage of conventional ray theory is that it does not include diffraction, including effects that describe the low-frequency dependence ("degree of trapping") of ducted propagation.

C. Wavenumber Representation or Spectral Solution

The wave equation can be solved efficiently with spectral methods when the ocean environment does not vary with range. The term "Fast Field Program (FFP)" had been used because the spectral methods became practical with

the advent of the fast Fourier transform (FFT). Assume a solution of Eq. (26) of the form

$$G(\mathbf{r}, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d^2\mathbf{k}\, \mathrm{g}(\mathbf{k}, z, z_s) \exp[i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r}_s)], \quad (30)$$

which then leads to the equation for the depth-dependent Green's function, $\mathrm{g}(\mathbf{k}, z, z_s)$,

$$\frac{d^2\mathrm{g}}{dz^2} + (K^2(z) - k^2)\mathrm{g} = -\frac{1}{2\pi}\delta(z - z_s). \quad (31)$$

Furthermore, we assume azimuthal symmetry, $kr > 2\pi$ and $\mathbf{r}_s = 0$ so that Eq. (30) reduces to

$$G(r, z) = \frac{\exp(-i\pi/4)}{(2\pi r)^{1/2}} \int_{-\infty}^{\infty} dk\, (k)^{1/2} \mathrm{g}(k, z, z_s) \exp(ikr). \quad (32)$$

This integral is then evaluated using the FFT algorithm. Although the method was initially labeled "fast field" it is fairly slow because of the time required to calculate the Green's functions (solve Eq. 31). However, it has advantages when one wishes to calculate the "near-field" region or to include shear wave effects in elastic media; it is also often used as a benchmark for other less exact techniques. With a great deal of additional computational effort, this method is extendable to range-dependent environments.

D. Normal Mode Model

Rather than solve Eq. (31) for each g for the complete set of k's (typically thousands of times), one can utilize a normal mode expansion of the form

$$g(\mathbf{k}, z) = \sum a_n(\mathbf{k}) u_n(z), \quad (33)$$

where the quantities u_n are eigenfunctions of the following eigenvalue problem:

$$\frac{d^2u_n}{dz^2} + \left[K^2(z) - k_n^2\right]u_n(z) = 0. \quad (34)$$

The eigenfunctions, u_n, are zero at $z = 0$, satisfy the local boundary conditions descriptive of the ocean-bottom properties, and satisfy a radiation condition for $z \to \infty$. They form an orthonormal set in a Hilbert space with weighting function $\rho(z)$, the local density. The range of discrete eigenvalues corresponding to the poles in the integrand of Eq. (32) is given by the condition

$$\min[K(z)] < k_n < \max[K(z)]. \quad (35)$$

These discrete eigenvalues correspond to discrete angles within the critical angle cone in Fig. 6a as discussed in Section I.B.3. The eigenvalues k_n typically have a small imaginary part α_n, which serves as the modal attenuation representative of all the losses in the ocean environment.

Solving Eq. (26) using the normal mode expansion given by Eq. (33) yields (for the source at the origin)

$$G(r, z) = \frac{i}{4}\rho(z_s) \sum_n u_n(z_s) u_n(z) H_0^1(k_n r). \quad (36)$$

The asymptotic form of the Hankel function can be used in the above equation to obtain the well-known normal mode representation of a cylindrical (axis is depth) waveguide:

$$G(r, z) = \frac{i\rho(z_s)}{(8\pi r)^{1/2}} \exp(-i\pi/4) \times \sum_n \frac{u_n(z_s)u_n(z)}{k_n^{1/2}} \exp(ik_n r). \quad (37)$$

Equation (37) is a far field solution of the wave equation and neglects the continuous spectrum ($k_n < \min[K(z)]$ of Ineq. 35) of modes. For purposes of illustrating the various portions of the acoustic field, we note that k_n is a horizontal wave number so that a "ray angle" associated with a mode with respect to the horizontal can be taken to be $\theta = \cos^{-1}[k_n/K(z)]$. For a simple waveguide, the maximum sound speed is the bottom sound speed corresponding to $\min[K(z)]$. At this value of $K(z)$, we have from Snell's law $\theta = \theta_c$, the bottom critical angle. In effect, if we look at a ray picture of the modes, the continuous portion of the mode spectrum corresponds to rays with grazing angles greater than the bottom critical angle of Fig. 8b and therefore outside the cone of Fig. 6a. This portion undergoes severe loss. Hence, we note that the continuous spectrum is the near (vertical) field and the discrete spectrum is the far (more horizontal, profile dependent) field falling within the cone in Fig. 6a.

The advantages of the normal mode procedure are that (1) the solution is available for all source and receiver configurations once the eigenvalue problem is solved; (2) it is easily extended to moderately range-dependent conditions using the adiabatic approximation; (3) it can be applied (with more effort) to extremely range-dependent environments using coupled mode theory. However, it does not include a full representation of the near field.

E. Adiabatic Mode Theory

All of the range-independent normal mode "machinery" developed for environmental ocean acoustic modeling applications can be adapted to mildly range-dependent conditions using adiabatic mode theory. The underlying assumption is that individual propagating normal modes adapt (but do not scatter or "couple" into each other) to the local environment. The coefficients of the mode expansion, a_n in Eq. (33), now become mild functions of range, i.e., $a_n(\mathbf{k}) \to a_n(\mathbf{k}, \mathbf{r})$. This modifies Eq. (32) as follows:

$$G(\mathbf{r}, z) = \frac{i\rho(z_s)}{(8\pi r)^{1/2}} \exp(-i\pi/4) \times \sum_n \frac{u_n(z_s)v_n(z)}{\overline{k_n}^{1/2}} \exp(i\overline{k_n}r). \quad (38)$$

where the range-averaged wavenumber (eigenvalue) is

$$\overline{k_n} = \frac{1}{r}\int_0^r k_n(r')\,dr', \quad (39)$$

and the $k_n(r')$ are obtained at each range segment from the eigenvalue problem Eq. (34) evaluated for the environment at that particular range along the path. The quantities u_n and v_n are the sets of modes at the source and the field positions, respectively.

Simply stated, the adiabatic mode theory leads to a description of sound propagation such that the acoustic field is a function of the modal structure at both the source and the receiver and some average propagation conditions between the two. Thus, for example, when sound emanates from a shallow region where only two discrete modes exist and propagates into a deeper region with the same bottom (same critical angle), the two modes from the shallow region adapt to the form of the first two modes in the deep region. However, the deep region can support many more modes; intuitively, we therefore expect the resulting two modes in the deep region will take up a smaller more horizontal part of the cone of Fig. 6a than they take up in the shallow region. This means that sound rays going from shallow to deep tend to become more horizontal, which is consistent with a ray picture of downslope propagation. Finally, fully coupled mode theory for range-dependent environments has been developed but requires extremely intensive computation.

1. Parabolic Equation Model (PE)

The PE method was introduced into ocean acoustics and made viable with the development of the "Tappert split-step algorithm" which utilized FFTs at each range step. Subsequent numerical developments greatly expanded the applicability of parabolic equation.

2. Standard PE Split–Step Algorithm

The PE method is presently the most practical and encompassing wave-theoretic range-dependent propagation model. In its simplest form, it is a far-field narrow-angle ($\sim \pm 20^\circ$ with respect to the horizontal—adequate for most underwater propagation problems) approximation to the wave equation. Assuming azimuthal symmetry about a source, we express the solution of Eq. (21) in cylindrical coordinates in a source free region in the form

$$p(r, z) = \psi(r, z) \cdot H(r), \quad (40)$$

and we define $K^2(r, z) \equiv K_0^2 n^2$, n therefore being an "index of refraction" c_0/c, where c_0 is a reference sound speed. Substituting Eq. (40) into Eq. (21) and taking K_0^2 as the separation constant, we end up with a Bessel equation for H which has a Hankel function as the outgoing solution. If we use the asymptotic form of the Hankel function, $H_0^1(K_0 r)$, and invoke the "paraxial" (narrow angle) approximation,

$$\frac{\partial^2\psi}{\partial r^2} \ll 2K_0\frac{\partial\psi}{\partial r}, \quad (41)$$

we obtain the parabolic equation (in r),

$$\frac{\partial^2\psi}{\partial z^2} + 2iK_0\frac{\partial\psi}{\partial r} + K_0^2(n^2 - 1)\psi = 0, \quad (42)$$

where we note that n is a function of range and depth. We use a marching solution to solve the parabolic equation. There has been an assortment of numerical solutions, but the one that still remains a standard is the so-called "split-step" range-marching algorithm,

$$\psi(r + \Delta r, z) = \exp\left[\frac{iK_0}{2}(n^2 - 1)\Delta r\right]\mathcal{F}^{-1} \times \left[\left(\exp\left(-\frac{i\Delta r}{2K_0}s^2\right)\right) \cdot \mathcal{F}[\psi(r, z)]\right], \quad (43)$$

which is often referred to as the "split-step" marching solution to the PE. The Fourier transforms $\mathcal{F}$ are performed using FFTs. Equation (43) is the solution for n constant, but the error introduced when n (profile or bathymetry) varies with range and depth can be made arbitrarily small by increasing the transform size and decreasing the range-step size. It is possible to modify split-step algorithm to increase its accuracy with respect to higher angle propagation.

3. Generalized or Higher-Order PE Methods

Methods of solving the parabolic equation, including extensions to higher angle propagation, elastic media, and direct time domain solutions including nonlinear effects, have recently appeared. In particular, accurate high angle solutions are important when the evironment supports acoustic paths that become more vertical such as when the bottom has a very high speed and hence, large critical angle with respect to the horizontal. In addition, for elastic propagation, the compressional and shear waves span a wide angle interval. Finally, Fourier synthesis for pulse modeling requires high accurate in phase and the high angle PE's are more accurate in phase, even at the low angles.

Equation (42) with the second-order range derivative which was neglected because of Ineq. (41) can be written in operator notation as

$$[P^2 + 2iK_0P + K_0^2(Q^2 - 1)]\psi = 0, \qquad (44)$$

where

$$P \equiv \frac{\partial}{\partial r}, \qquad Q \equiv \sqrt{n^2 + \frac{1}{K_0^2}\frac{\partial^2}{\partial z^2}}. \qquad (45)$$

Factoring Eq. (45) assuming weak range dependence and retaining only the factor associated with outgoing propagation yields a one-way equation

$$P\psi = iK_0(Q - 1)\psi \qquad (46)$$

which is a generalization of the parabolic equation beyond the narrow angle approximation associated with Ineq. (38). If we define $Q = \sqrt{1+q}$ and expand Q in a Taylor series as a function of q, the standard PE method is recovered by $Q \approx 1 + 0.5q$. The wide-angle PE to arbitrary accuracy in angle, phase, etc, can be obtained from a Padé series representation of the Q operator,

$$Q \equiv \sqrt{1+q} = 1 + \sum_{j=1}^{n} \frac{a_{j,n}q}{1 + b_{j,n}q} + \mathcal{O}(q^{2n+1}), \qquad (47)$$

where n is the number of terms in the Padé expansion and

$$a_{j,n} = \frac{2}{2n+1}\sin^2\left(\frac{j\pi}{2n+1}\right), \quad b_{j,n} = \cos^2\left(\frac{j\pi}{2n+1}\right). \qquad (48)$$

The solution of Eq. (46) using Eqs. (47) and (48) has been implemented using finite difference techniques for fluid and elastic media.

V. QUANTITATIVE DESCRIPTION OF PROPAGATION

All of the models described above attempt to describe reality and to solve in one way or another the Helmholtz equation. They therefore should be consistent, and there is much insight to be gained from understanding this consistency. The models ultimately compute propagation loss which is taken as the decibel ratio (see Appendix) of the pressure at the field point to a reference pressure, typically one meter from the source.

Figure 13 shows convergence zone type propagation for a simplified profile. The ray trace in Fig. 13b shows the cyclic focusing discussed in Section I.B. The same profile is used to calculate normal modes, shown in Fig. 13c, which when summed according to Eq. (37) exhibit the same cyclic pattern as the ray picture. Figure 13d shows both the normal mode (wave theory) and ray theory result. Ray theory exhibits sharply bounded shadow regions

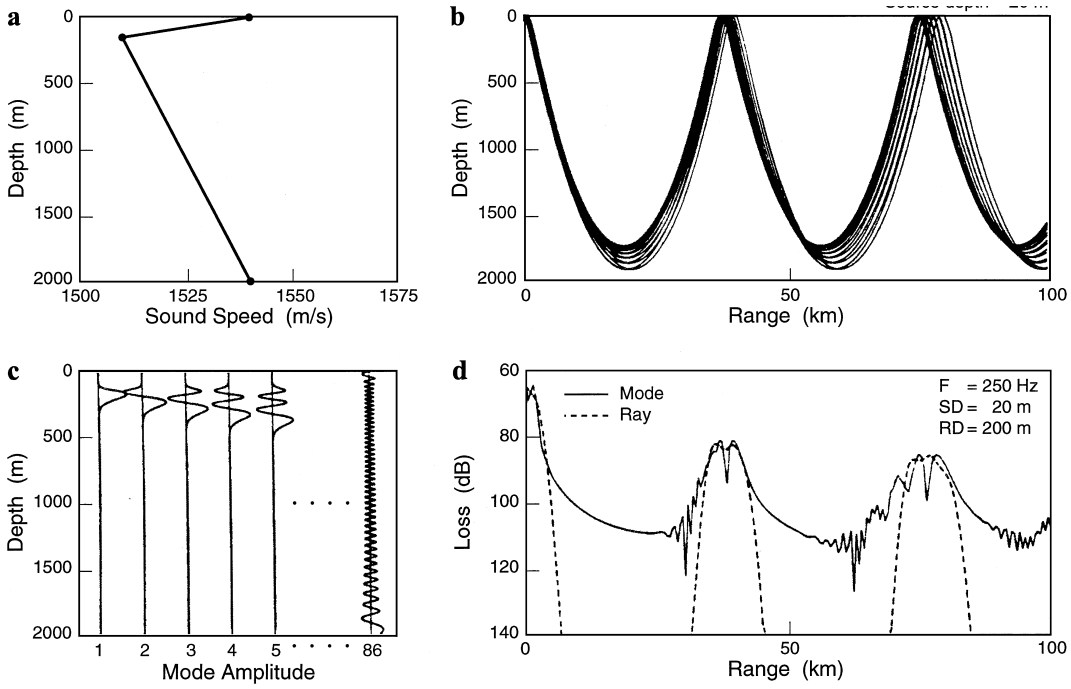


FIGURE 13 Consistency between ray theory and normal mode theory. (a) Sound speed profile. (b) Ray trace. (c) Normal modes. (d) Propagation calculations.

as expected, whereas the normal mode theory, which includes diffraction, shows that the acoustic field does exist in the shadow regions, and the convergence zones have structure.

Normal mode models sum the discrete modes which roughly correspond to angles of propagation within the cone of Fig. 6a. The spectral method can include the full field, discrete plus continuous, the latter corresponding to larger angles. The discussion following Eq. (37) defines these angles in terms of horizontal wavenumbers, and eigenvalues of the normal mode problem are a discrete set of horizontal wavenumbers. Hence the integrand (Green's function) of the spectral method has peaks at the eigenvalues associated with the normal modes. These peeks are shown on the right of Fig. 14a. The smoother portion of the spectrum is the continuous part corresponding to the larger angles. Therefore, the consistency we expect between the normal mode and the spectral method and the physics of Fig. 6 is that the continuous portion of the spectral solution decays rapidly with range so that there should be complete agreement at long ranges between normal mode and spectral solutions. The Lloyd mirror effect, a near-field effect, should also be exhibited in the spectral solution but not the normal mode solution. These aspects are apparent in Fig. 14b. The PE solution is in good agreement with the other solutions but with some phase error associated with the average wavenumber that must be chosen in the split-step method. The PE solution, which contains part of the continuous spectrum including the Lloyd mirror beams, is more accurate than the normal mode solution at short range; however, the generalized PE can be made arbitrarily accurate at short range by including more expansion terms in Eq. (47).

Range-dependent results are shown in Fig. 15. A ray trace, a ray trace field result, a PE result, and data are plotted together for a range-dependent sound speed profile environment. The models agree with the data in general, with the exception that the ray results predict too sharp a leading edge of the convergence zone.

Upslope propagation is modeled with the PE in Fig. 16. As the field propagates upslope, sound is dumped into the bottom in what appears to be discrete beams. The flat region has three modes and each is cut off successively as sound propagates into shallower water. The ray picture also has a consistent explanation of this phenomenon. The rays for each mode become steeper as they propagate upslope. When the ray angle exceeds the critical angle, the

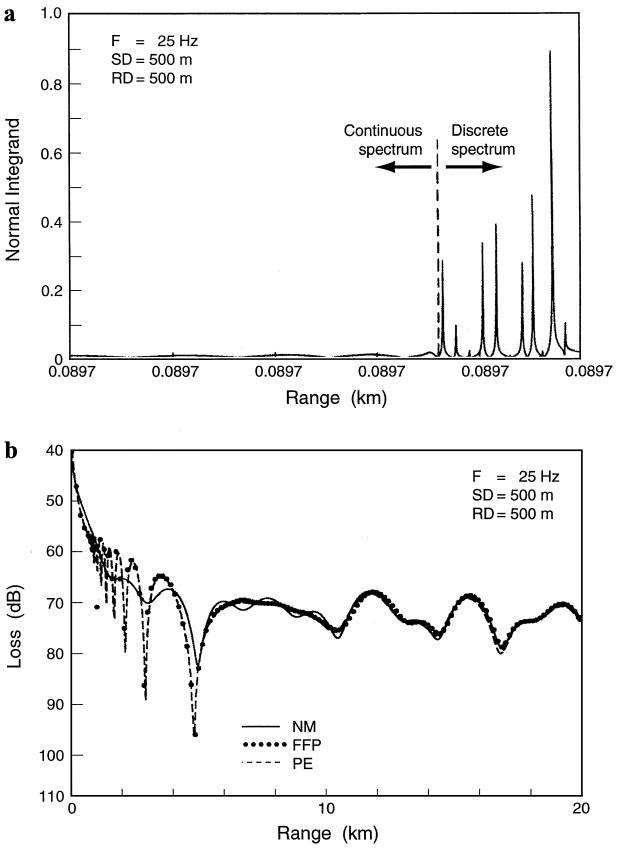


FIGURE 14 Relationship between FFP, NM, and PE computations. (a) FFP Green's function from Eq. (31). (b) Normal mode, spectral (FFP), and PE propagation results showing some agreement in near field and complete agreement in far field.

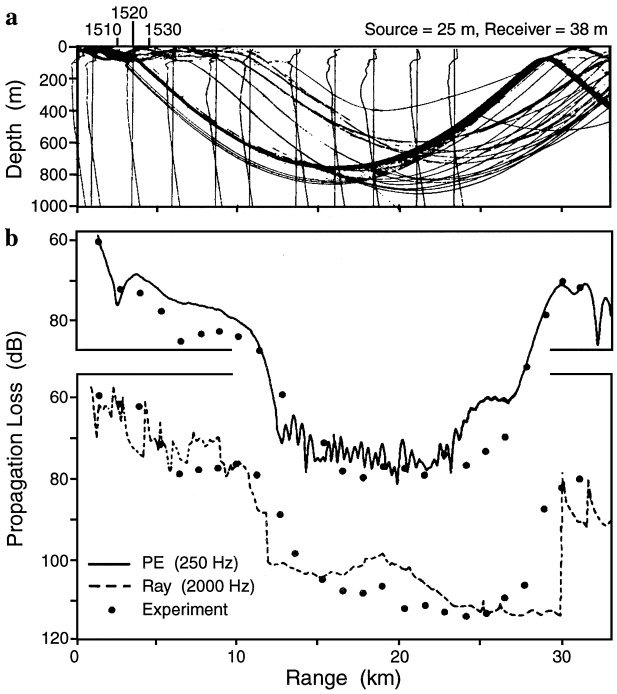


FIGURE 15 Model and data comparison for a range dependent case. (a) Profiles and ray trace for a case of a surface duct disappearing. (b) 250 Hz PE and 2 kHz Ray trace comparisons with data.

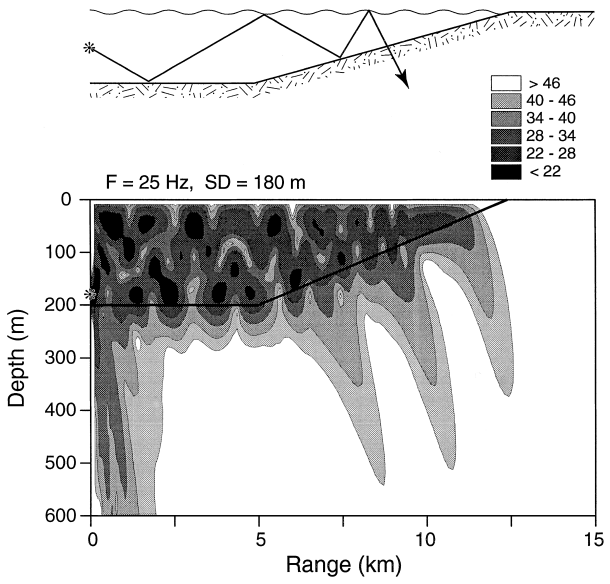


FIGURE 16 Relation between upslope propagation (from PE calculation) showing individual mode cutoff and energy dumping in the bottom, and a corresponding ray schematic.

sound is significantly transmitted into the bottom. The locations where this takes place for each of the modes are identified by the three arrows.

As a final example of how physical insight can be derived from models, we present a range-independent normal mode study of the optimum frequency of propagation in a shallow water environment with summer and winter (dashed lines) profiles as indicated in Fig. 17a with the source (S) and receiver (R) also indicated. Frequency versus range contours of propagation loss obtained from a wideband experiment (analyzed in 1/3 octave bands) are shown in Fig. 17b. One obtains the conventional type of propagation loss curves by taking a horizontal cut through the contour in Fig. 17b, with the result shown in Fig. 17c. Figure 17d is a model result from an incoherent (no cross terms) sum of normal modes. We note here, as an aside, that in shallow water environments, propagation loss obtained by incoherently summing the modes is approximately equal to 1/3 octave frequency averaging, which has the effect of averaging away modal interference. The frequency versus range contours reveal an optimum frequency in the 200–400 Hz region. This can be seen by observing the 80 dB contour which goes out to long ranges in the region, whereas other frequencies, at say a range of 70 km, have much higher losses.

VI. SONAR ARRAY PROCESSING

Temporal processing such as digital signal processing is common to many fields. In this section we emphasize applications to underwater acoustics, mainly concentrating on spatial processing. Further, the array processing discussed below for passive sonars is also applicable to active sonar signal processing. Spatial sampling of a sound field is usually done by an array of transducers, although the synthetic aperture array, in which a sensor is moved through space to obtain measurements in both the time and space domains, is an important exception. Spatial sampling is analogous to temporal sampling, with the sampling interval replaced by the sensor spacing vectors. The Nyquist criterion requires that the sensor spacing be at least twice the spatial wavenumber of the measured sound field.

A. Linear Plane Wave Beamforming and Spatiotemporal Sampling

The simplest example of array processing is phase shading in the frequency domain (or time delay in the time domain) to search for the bearing of a plane wave signal. This procedure is referred to as plane wave beamforming, or delay and sum beamforming in the time domain. For simplicity we consider a linear array, and we take θ as the bearing angle associated with the plane wave signal as shown in Fig. 18.

1. Frequency Domain Processing

A plane wave can be represented as

$$s(\theta) = \exp(i\mathbf{k}\cdot\mathbf{r}), \tag{49}$$

where we have suppressed the time dependence $\exp(-i\omega t)]$ and $k = |\mathbf{k}| = \omega/c$. The field is summed in phase if the receiving element (hydrophone or microphone) inputs at position $\mathbf{d_i}$ are multiplied by the complex conjugate of the plane wave phase factor,

$$w_i^* = \exp(-i\mathbf{k}\cdot\mathbf{d_i}) = \exp[-id(k\sin\theta_s)], \tag{50}$$

where θ_s is a scanning angle. This process will have a maximum when the scanning angle equals the incident angle of the signal.

The output of this beamforming process is denoted $B(\theta_s)$, but often it is the power ouput of the beamformer that is of interest:

$$|B(\theta_s)|^2 = \left|\sum_{i=1}^{m} w_i^*(\theta_s)[s_i(\theta) + n_i]\right|^2$$

$$= \sum_{i,j=1}^{m} w_i^*(\theta_s)(s_{ij} + n_{ij})w_j(\theta_s), \tag{51}$$

where s_i and n_i are the signal and noise at the ith receiving element and where $s_{ij} + n_{ij}$ are elements of a cross-spectral density matrix which, when obtained from data, would involve Fourier transforms and ensemble averages

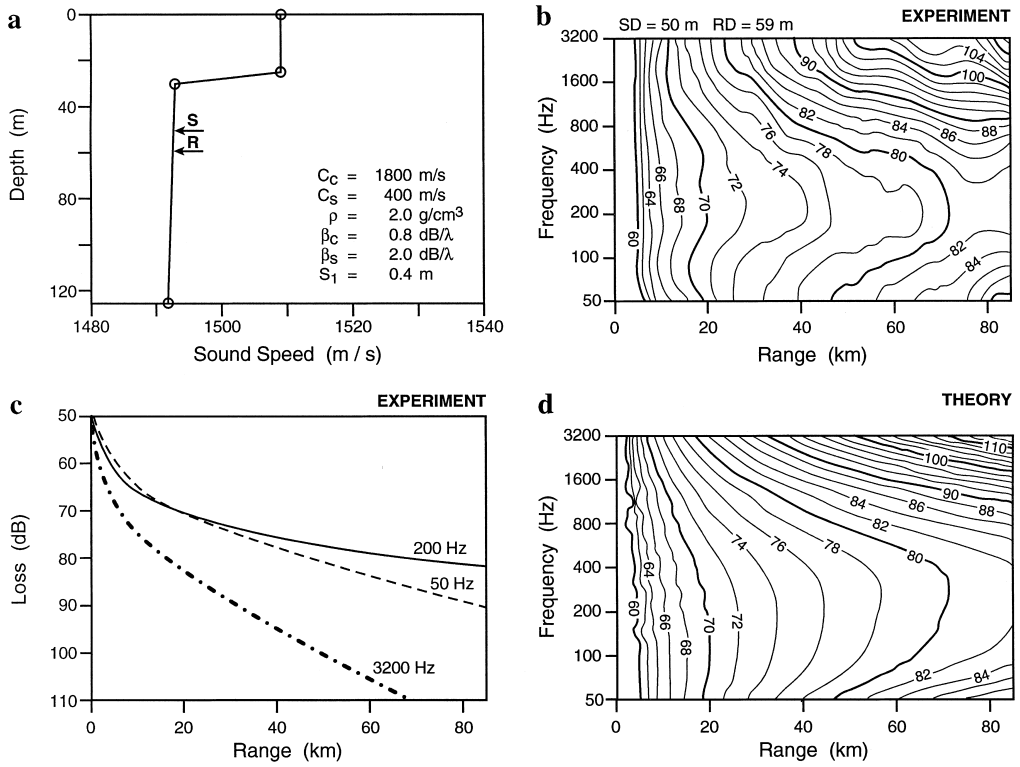


FIGURE 17 (a) Shallow water environment with summer sound speed profile. (b) Frequency versus range contours of propagation loss obtained from a wide-band experiment (analyzed in 1/3 octave bands). (c) Propagation curves at three frequencies corresponding from experiment contour curves. (d) Theoretical result using normal mode model.

as mentioned in the introduction and in the discussion following Eq. (53) augmented by Fig. 19. In writing down the right-hand side of Eq. (51), the signal and noise fields were assumed to be mutually incoherent.

We can write the above expression in matrix notation where the boldface lower-case letters denote vectors and boldface upper-case letters denote matrices. Define a steering column vector $\mathbf{w}$ whose ith element is w_i and a cross-spectral density matrix (CSDM) $\mathbf{K}$ of the signal and noise with elements $K_{ij} = s_{ij} + n_{ij}$ since the signal and noise are assumed to be independent. Equation (51) can be rewritten as

$$|B(\theta_s)|^2 = \mathbf{w}^\dagger(\theta_s)\mathbf{K}(\theta_{true})\mathbf{w}(\theta_s) \equiv \mathbf{w}^\dagger\mathbf{K}\mathbf{w} \qquad (52)$$

where "†" denotes the complex transpose operation. The CSDM or the covariance of the field is composed of uncorrelated signal and noise covariances,

$$\mathbf{K} = \mathbf{K_s} + \mathbf{K_n}. \qquad (53)$$

The data across the array as represented in the matrix $\mathbf{K}$ contain the information that the source is in the direction θ_{true}. Sometimes $\mathbf{w}(\theta_s)$ is referred to as a replica, and the above beamforming process is viewed as matching the received data across the array with a replica. The type of

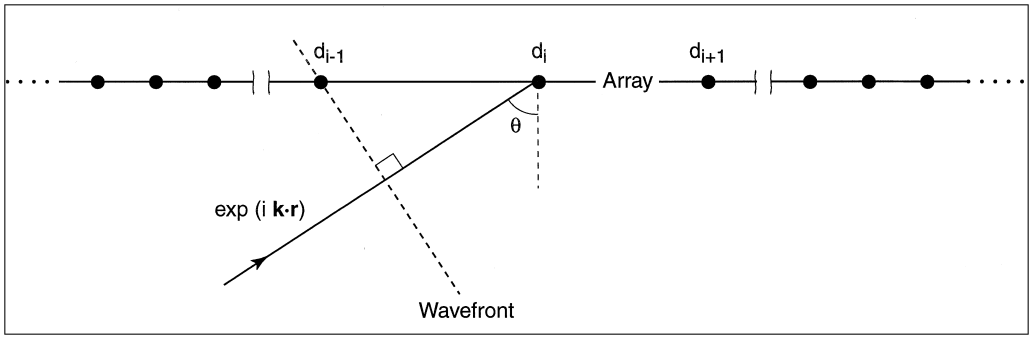


FIGURE 18 Geometry for plane wave beamforming.

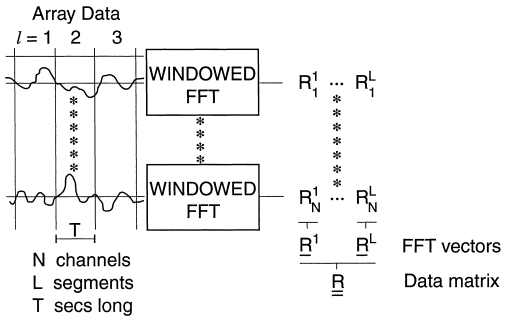


FIGURE 19 Array, narrowband model, and sample covariance matrix estimation.

beamformer represented by Eq. (52) is called a linear or a Bartlett beamformer.

For the sample covariance estimation we assume we have an array with N sensors located at $\mathbf{d}_i$, $i = 1, N$, and a narrowband model as illustrated in Fig. 19. These covariances are estimated by segmenting the received data, $r_i(t)$ into "snapshots" using a sampling window, $W(t)$, that is unity in the interval

$$[0, T\omega], R_i^l(f) = \int_{T_l}^{T_l+T_w} r_i(t)W(t-T_l)e^{-j2\pi ft}\,dt \quad (54)$$

where, here, the notation uses frequency, f, rather than angular frequency, ω. In most beamforming algorithms the data vectors are averaged to form the sample covariance matrix

$$\hat{\mathbf{K}}(f) = \frac{1}{L}\sum_{l=1}^{L}\mathbf{R}^l(f)\mathbf{R}^l(f)^H \quad (55)$$

where L is the number of snapshots.

2. Time Domain Processing

Time delay is the time domain analogy to phase shading in the frequency domain. This can be derived formally by taking the Fourier transform of the beamforming process with the result that the beamformer output is

$$B(t) = \sum_i r_i\left(t - \frac{d_i}{c}\sin\theta\right), \quad (56)$$

where $r_i(\cdot)$ is the time domain data at the ith phone. This process is referred to as delay and sum beamforming; the delay is simply the time interval associated with the phase shift in the frequency domain array processing.

B. Some Beamformer Properties

Figure 20 shows the output results of some plane wave beamformers for the cases of one and two incident signals. To be noted are the sidelobes of the Bartlett processor and

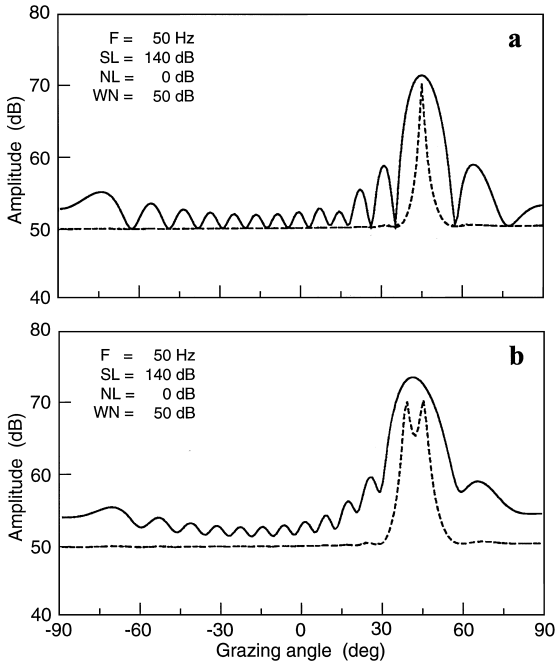


FIGURE 20 Beamformer outputs. (a) Single sources at a bearing of 45°. (b) Two sources with 6.3° angular separation. Solid line: linear processor (Bartlett). Dashed line: minimum variance distortionless processor (MV).

the high-resolution performance of the adaptive processors (discussed in the next section). Some of the general attributes of an array processor are:

- The main response axis (MRA): Generally, one normalizes the beampattern to have 0 dB, or unity gain in the steered direction.
- Beamwidth: An array with finite extent, or aperture, must have a finite beamwidth centered about the MRA which is termed the "main lobe."
- Sidelobes: Sidelobes are angular or wavenumber regions where the array has a relatively strong response. Sometimes they can be comparable to the MRA, but in a well-designed array, they are usually −20 dB or lower, i.e, the response of the array is less than 0.1 to a signal in the direction of a sidelobe.
- Wavenumber processing: Rather than scan through incident angles, one can scan through wavenumbers, $k\sin\theta_s \equiv \kappa_s$; scanning through all possible values of κ_s results in nonphysical angles which correspond to waves not propagating at the acoustic medium speed. Such waves can exist, such as those associated with array vibrations. The beams associated with these wavenumbers are sometimes referred to as virtual beams. An important aspect of these beams is that their sidelobes can be in the physical angle region thereby interfering with acoustic propagating signals.

- Array gain: The array gain is defined as the decibel ratio of the signal-to-noise ratios of the array output to a single phone output. If the noise field is isotropic, the array gain is also termed the directivity index.

C. Adaptive Processing

There are high-resolution methods to suppress sidelobes, usually referred to as adaptive methods since the signal processing procedure constructs weight vectors that depend on the received data itself. We briefly describe one of these procedures: the Minimum Variance Distortionless Processor (MVDP), sometimes also called the Maximum Likelihood Method (MLM) directional spectrum estimation procedure.

We seek a weight vector $\mathbf{w}_{MV}$ applied to the matrix $\mathbf{K}$ such that its effect will be to minimize the output of the beamformer, Eq. (52), except in the look direction where we want the signal to pass through undistorted. The weight vector is therefore chosen to minimize the functional

$$F = \mathbf{w}_{MV}^{\dagger}\mathbf{K}\mathbf{w}_{MV} + \alpha\left(\mathbf{w}_{MV}^{\dagger}\mathbf{w} - 1\right). \tag{57}$$

The first term is the mean-square output of the array and the second term incorporates the constraint of unity gain by means of the Lagrangian multiplier α. Following the method of Lagrange multipliers, we obtain the MV weight vector,

$$\mathbf{w}_{MV} = \frac{\mathbf{K}^{-1}\mathbf{w}}{\mathbf{w}^{\dagger}\mathbf{K}^{-1}\mathbf{w}}. \tag{58}$$

This new weight vector depends on the received data as represented by the cross-spectral density matrix; hence, the method is "adaptive." Substituting back into the quadratic form of Eq. (52) gives the output of our MV processor,

$$B_{MV}(\theta_s) = \left[\mathbf{w}^{\dagger}(\theta_s)\mathbf{K}^{-1}(\theta_{true})\mathbf{w}(\theta_s)\right]^{-1}. \tag{59}$$

The MV beamformer should have the same peak value at θ_{true} as the Bartlett beamformer, Eq. (52), but with sidelobes suppressed and narrower main beam, indicating that it is a high-resolution beamformer. Examples are shown in Fig. 20.

D. Matched Field Processing

Matched field processing (MFP) is the three-dimensional generalization of the conventional lower-dimensional plane wave beamformer that matches the measured field at the array with replicas of the expected field for all source locations. These replicas, $\mathbf{w}$, are derived from propagation models as discussed in Section IV. The unique spatial structure of the field permits localization in range, depth, and azimuth depending on the array geometry and complexity of the ocean environment. The process is shown schematically in Fig. 21. MFP consists of systematically placing a test point source at each point of a search grid, computing the acoustic field (replicas) at all the elements of the array, and then correlating this modeled field with the data from the real point source, $\mathbf{K}(\mathbf{a}_{true})$, whose location is unknown. When the test point source is colocated with the true point source, the correlation will be a maximum. The output of this matched field processor, denoted $S(\mathbf{a})$ to indicate the generalization beyond plane wave beamforming, at each point in space $\mathbf{a}$ is, in analogy to Eq. (52),

$$S(\mathbf{a}) = \mathbf{w}^{\dagger}(\mathbf{a})\mathbf{K}(\mathbf{a}_{true})\mathbf{w}(\mathbf{a}), \tag{60}$$

where the peak of the output of the beamformer, $S(\mathbf{a})$, is at $\mathbf{a}_{true}$. $S(\mathbf{a})$ is also referred to as the ambiguity function (or surface) of the matched field processor because it also contains ambiguous peaks which are analogous to the sidelobes of a conventional plane wave beamformer. Sidelobe suppression can often be accomplished by using a nonlinear beamformer such as the MLM beamformer:

$$S_{MV}(\mathbf{a}) = \left[\mathbf{w}^{\dagger}(\mathbf{a})\mathbf{K}^{-1}(\mathbf{a}_{true})\mathbf{w}(\mathbf{a})\right]^{-1}. \tag{61}$$

A simulation vertical receive array example of the Bartlett and MVDP MFP processors for an ocean acoustic waveguide with a high signal-to-noise ratio is shown in Fig. 22. The two main factors that limit performance of MFP are noise and the ability to accurately model the environment. Related to MFP is matched field tomography (MFT) searches for the environmental parameters controlling the propagation (for example, the index of refraction which may be a spatially dependent coefficient of the wave equation) rather than source location.

VII. ACTIVE SONAR PROCESSING

An active sonar system transmits a pulse and extracts information from the echo it receives as opposed to a passive sonar system which extracts information from signals received from radiating sources. An active sonar system and its associated waveform is designed to detect targets and estimate their range, Doppler (speed), and bearing or to determine some properties of the medium such as ocean-bottom bathymetry, ocean currents, winds, particulate concentration, etc. The spatial processing methods already discussed are applicable to the active problem, so that in this section we emphasize the temporal aspects of active signal processing.

A. Active Sonar Signal Processing

The basic elements of an active sonar are: the (waveform) transmitter, the channel through which the signal, echo, and interference propagate, and the receiver. The receiver

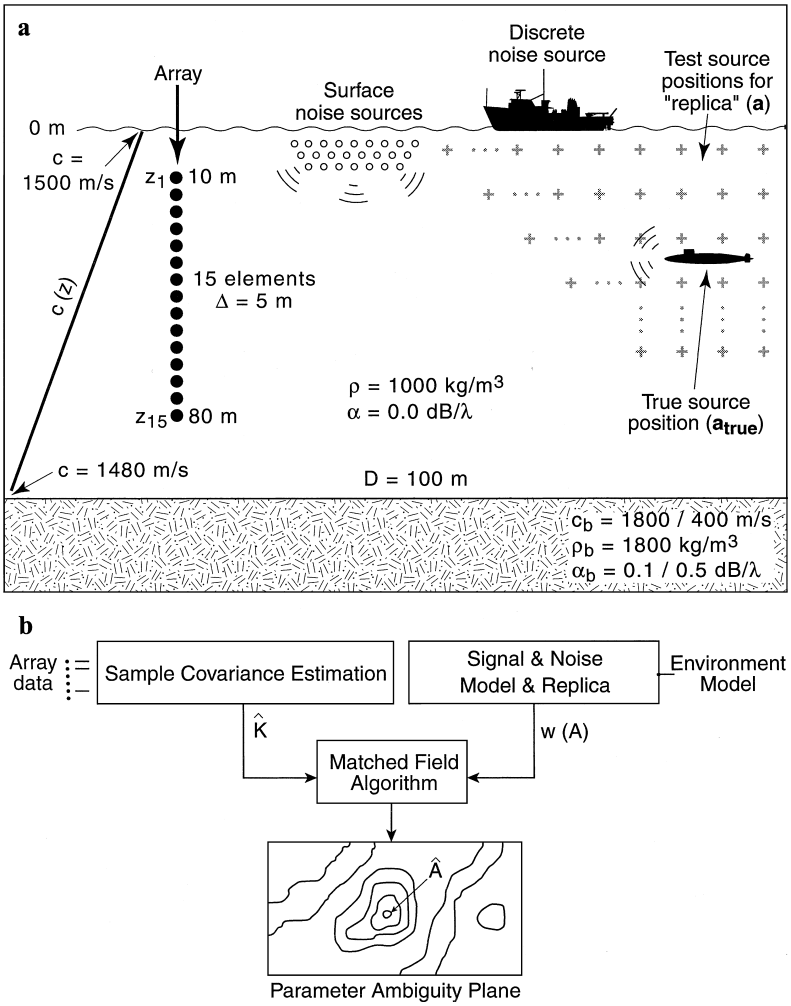


FIGURE 21 Matched field processing. (a) The true source location is obtained by modeling data at an array from a set of grid points and comparing the model with the actual data on the array. (b) Schematic diagram of the matched field processor.

consists of some sort of matched filter, a square law device, and possibly a threshold device for the detector and range, Doppler, and bearing scanners for the esimator.

The matched filter maximizes the ratio of the peak output signal power to the variance of the noise and is implemented by correlating the received signal with the transmitted signal. A simple description of the received signal, $r(t)$, is that it is an attenuated, delayed, and Doppler shifted version of the transmitted signal, $s_t(t)$,

$$r(t) \approx Re\left[\alpha e^{i\theta} \tilde{s}_t(t-T) e^{2\pi i f_c t} e^{2\pi i f_d t} + n(t)\right], \qquad (62)$$

where α is the attenuation transmission loss and target cross section, θ is a random phase from the range uncertainty compared to a wavelength, T is the range delay time, f_c is the center frequency of the transmitted signal, and f_d is the Doppler shift caused by the target. The correlation process will have an output related to the following process,

$$C(a) = \left| \int \tilde{r}(t) \tilde{s}(t;a)\, dt \right|^2 \qquad (63)$$

where $\tilde{s}(t;a)$ is a replica of the transmitted signal modified by a parameter set a which include the propagation-reflection process, e.g., range delay and Doppler rate. For the detection problem, the correlation receiver is used to generate a sufficient statistic which is the basis for a threshold comparison in making a decision if a target is present. The performance of the detector is described by receiving operating characteristic (ROC) curves which plot the detection of probability versus false alarm probability as parameterized by a statistic of the signal and noise levels. The parameters a set the range and Doppler value in the particular resolution cell of concern. To estimate

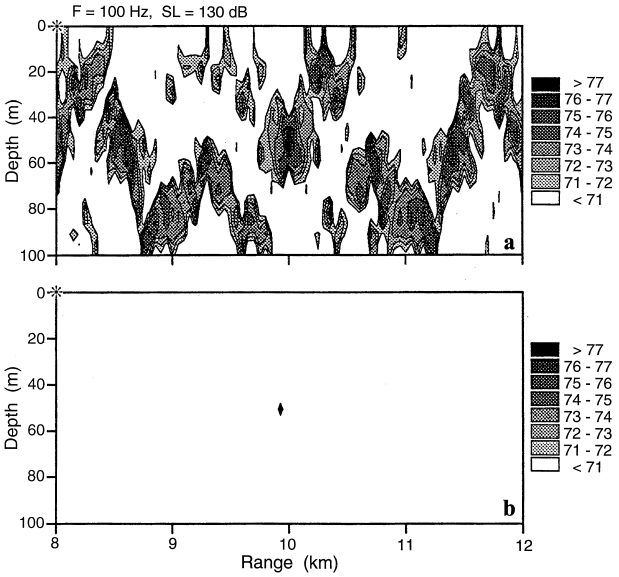


FIGURE 22 Simulated matched field results for the environment in Fig. 21a. (a) Bartlett MFP ambiguity surface. (b) Minimum variance distortionless MFP ambiguity surface.

these parameters, the correlation is done as a function of a.

For a matched filter operating in a background of white noise and detecting a point target in a given range-Doppler resolution cell, the detection signal-to-noise ratio depends on the average energy-to-noise ratio and not on the shape of the signal. The waveform becomes a factor when there is a reverberant environment and when one is concerned with estimating target range and Doppler. A waveform's potential for range and Doppler resolution can be ascertained from the ambiguity function of the transmitted signal. This ambiguity function is related to the correlation process of Eq. (63) for a transmitted signal scanned as a functions range and Doppler,

$$\Theta\left(\hat{T}, T_t, \hat{f}_d, f_{d_t}\right) \propto \left| \int \tilde{s}_t(t - T_t)\tilde{s}_t(t - \hat{T})\, e^{2\pi i (f_{d_t} - \hat{f}_d)t} /, dt \right|^2 \quad (64)$$

where T_t, f_{d_t} are the true target range (time) and Doppler and $\hat{T}$, $\hat{f}_d$ are the scanning estimates of range and Doppler. Figure 23 are sketches of ambiguities functions of some typical waveforms. The range resolution is determined by the reciprocal of the bandwidth and the Doppler resolution by the reciprocal of the duration. The coded or PR (pseudo-random) sequence can attain good resolution of both by appearing as long duration noise with a wide bandwidth. Ambiguity functions can be used to design desirable waveforms for particular situations. However, one must also consider the randomizing effect of the real

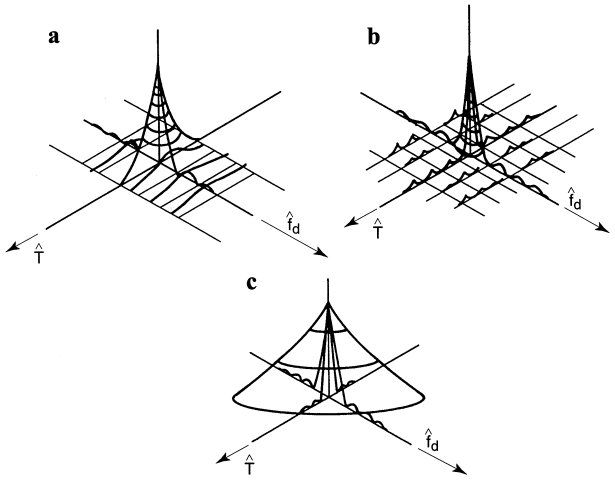


FIGURE 23 Ambiguity function for several sonar signals: (a) rectangular pulse; (b) coded pulses; (c) chirped FM pulse.

ocean. The *scattering function* describes how a transmitted signal statistically redistributes its energy in the reverberant ocean environment which causes multipath and Doppler spread. In particular, in a reverberation limited environment, increasing transmitted power only does not change the signal to reverberation level. Signal design should minimize the overlap of the ambiguity function of the target displaced to its range and Doppler and the scattering function.

B. Comparison of Processing for Detection, Communications, and Seabed Mapping

An underwater acoustic communication system is an active sonar system utilizing a channel which has many multipaths. In the communication problem, different signals, using some form of frequency or phase shift algorithm, are transmitted depending upon the message, and one wants to identify all the paths that were excited by the signal. In the detection problem, the same signal is transmitted and, and one wants to identify just those (range-Doppler) cells that are associated with the reflected energy of the target.

Furthermore, sequences of active signals can be used in both the target detection and the communication problem. For detecting a target, one wants to produce a track by smoothing a sequence of range-Doppler estimates. In communication systems, messages are often encoded into a sequence of transmissions for reliability and cryptographic concealment. For telemetry, equalization for compensating for mulitpath effects is often done without modeling the channel, but rather by using a reference signal or predefined sequence of symbols. Time varying intersymbol interference must be dealt with by some synchronization technique.

Maps and charts of the seabed are generated by plotting estimates of depths obtained by a sequence of reflections from the seabed. A single sounding does not convey an appreciable amount of information; a grid of sounding to determine the topographic relief is usually required. This necessitates a variety of processing methods; some simply compensate for the finite beamwidth of the sounding system, others compile the grid of points, interpolate the data, and contour the relief.

C. Travel Time Tomography

Tomography generally refers to applying some form of inverse theory to observations in order to infer properties of the propagation medium. The received field from a source emitting a pulse will be spread in time as a result of multipath structures in which different paths have different arrival times (or group speeds). Hence the arrival times are related to the acoustic sampling of the medium. In contrast, medical tomography utilizes the different attenuation of the paths rather than arrival time for the inversion process. Since the sound speed of the ocean is a function of temperature and other oceanographic parameters, the arrival structure can ultimately be related to a map of these oceanographic parameters. In its most primitive form, the inversion can be described by an algorithm discretizing the ocean into cells, computing travel times from candidate acoustic paths through these cells, and solving a set of equations which equate these travel times to the measured travel times. Each cell is characterized by an unknown sound speed. Typically, some baseline oceanographic information is known so that one searches for departures from this baseline information. Tomographic experiments have been performed to greater than megameter ranges. Pulse compression methods using sequences of signals are often employed in ocean tomographic experiments to enhance bandwidth and received signal strength.

VIII. APPENDIX: UNITS

The decibel (dB) is the dominant unit in underwater acoustics and denotes a ratio of intensities (not pressures) expressed in terms of a logarithmic (base 10) scale. Two intensities, I_1 and I_2, have a ratio, I_1/I_2, in decibels of $10 \log I_1/I_2$ dB. Absolute intensities can therefore be expressed by using a reference intensity. The presently accepted reference intensity is based on a reference pressure of one micropascal (μPa): the intensity of a plane wave having an rms pressurex equal to 10^{-5} dynes per square centimeter. Therefore, taking 1 μPa as I_2, a sound wave having an intensity, of, say, one million times that of a plane wave of rms pressure 1 μPa has a level of $10 \log(10^6/1) \equiv 60$ dB re 1 μPa. Pressure (p) ratios are expressed in dB re 1 μPa by taking $20 \log p_1/p_2$ where it is understood that the reference originates from the intensity of a plane wave of pressure equal to 1 μPa.

The average intensity, I, of a plane wave with rms pressure p in a medium of density ρ and sound speed c is $I = p^2/\rho c$. In seawater, ρc is 1.5×10^5 g cm^{-2} s^{-1} so that a plane wave of rms pressure 1 dyne/cm^2 has an intensity of 0.67×10^{-12} W/cm^2. Substituting the value of a micropascal for the rms pressure in the plane wave intensity expression, we find that a plane wave pressure of 1 μPa corresponds to an intensity of 0.67×10^{-22} W/cm^2 (i.e., 0 dB re 1 μPa).

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICAL MEASUREMENT • LIQUIDS, STRUCTURE AND DYNAMICS • PHYSICAL OCEANOGRAPHY • SIGNAL PROCESSING, ACOUSTIC • SIGNAL PROCESSING, DIGITAL • SIGNAL PROCESSING, GENERAL • WAVE PHENOMENA

BIBLIOGRAPHY

Baggeroer, A. B. (1978). In *Applications of Digital Signal Processing* (A. V. Oppenheim, ed.), Prentice Hall, Englewood Cliffs, NJ.

Brekhovskikh, L. M., and Lysanov, Y. P. (1991). *Fundamentals of Ocean Acoustics*, Springer-Verlag, Berlin.

Burdic, W. S. (1984). *Underwater Acoustic Signal Analysis*, Prentice Hall, Englewood Cliffs, NJ.

Collins, M. D., and Siegmann, W. L. (2001). *Parabolic Wave Equations with Applications,* Springer-AIP, New York.

deMoustier, C. *Int. Hydrogr. Rev*. **LXV,** 25.

Jensen, F. B., Kuperman, W. A., Porter, M. B., and Schmidt, H. (1994). *Computational Ocean Acoustics*, AIP Press, New York.

Johnson, D. H., and Dudgeon, D. E. (1993). *Array Signal Processing: Concepts and Techniques*, PTR Prentice Hall, Englewood Cliffs.

Keller, J. B., and Papadakis, J. S., eds. (1977). *Wave Propagation in Underwater Acoustics*, Springer-Verlag, New York.

Medwin, H., and Clay, C. S. (1997). *Fundamentals of Acoustical Oceanography*, Academic Press, San Diego.

Munk, W., Worcester, P., and Wunsch, C. (1995). *Acoustic Tomography*, Cambridge Univ. Press, Cambridge.

Proakis, J. G. (1989). *Digital Communications,* McGraw Hill, New York.

Ross, D. (1976). *Mechanics of Underwater Noise*, Pergamon, New York.

Ogilvy, J. A. (1987). Wave Scattering from Rough Surfaces. *Rep. Prog. Phys*. **50,** 1553–1608.

Urick, R. J. (1983). *Principles of Underwater Sound*, McGraw Hill, New York.

VanTrees, H. L. (1971). *Detection Estimation and Modulation Theory*, Wiley, New York.

Wilson, O. B. (1985). *An Introduction to the Theory and Design of Sonar Transducers*, U.S. GPO, Washington, DC.



Vibration, Mechanical

Marie Dillon Dahleh
University of California, Santa Barbara

William T. Thomson
University of California, Santa Barbara, deceased

I. Definitions
II. Systems with One Degree of Freedom
III. Systems with Multiple Degrees of Freedom
IV. Continuous Systems: Normal Modes
V. Lagrange's Equation: Generalized Coordinates
VI. Approximate and Numerical Methods
VII. Conclusions

GLOSSARY

Characteristic equation Algebraic equation from which natural frequencies are calculated.
Circular frequency Frequency measured in radians per second.
Eigenvalue Quantity associated with natural frequencies.
Eigenvector Vector column of natural modes.
Free vibration Vibration in the absence of external excitation.
Modal matrix Matrix of eigenvectors.
Mode Deflection shape.
Node Point of zero deflection.
Shock spectrum Plot of maximum peak response versus period in terms of pulse time.
Transient vibration Vibration due to shock.
Viscous damping Damping proportional to velocity.

VIBRATION is a back-and-forth oscillation about an equilibrium position with a wavelike character of periodicity. For example, the swinging motion of a pendulum is a vibration that is visible to the eye. It has a period of oscillation that can be measured by a stopwatch, and its amplitude of oscillation can be observed from the extreme excursion of the swing. The number of oscillations in a unit of time is called the frequency. More often, the vibration is so small in amplitude that it is not observable by the naked eye and its motion must be measured by a sensitive instrument. Sound is a vibration of the air, and its oscillation may be quite complicated with its wave profile containing many frequencies. Everything in nature vibrates and has a rate of vibration. Vibration is a universal phenomenon since all bodies possessing mass and elasticity are capable of it.

I. DEFINITIONS

Mass and elasticity are the elements of a vibrating system. Their distribution establishes the natural frequencies of the system. The mass in motion possesses kinetic energy, whereas potential energy is stored in the deformation of the elastic element. In a conservative system a continuous

interchange of kinetic and potential energy takes place under constant total energy. Actually, some dissipation of energy into heat or sound is encountered, diminishing the motion of the system. Such action is represented by a damper, the third element of the system. Thus, the basic dynamic model of a simple vibratory system consists of a mass, a massless spring, and a damper.

Oscillatory motions are generally periodic: that is, they repeat themselves in equal intervals of time called the period τ. The motion completed during the period is referred to as the cycle. The number of complete cycles of motion in a unit of time is called the frequency of vibration f, which is the reciprocal of the period, $f = 1/\tau$.

Vibrations fall into two general classes: free and forced. Free vibration takes place when a system vibrates under the action of forces inherent in the system itself and when external impressed forces are absent. A system under free vibration vibrates at one or more of its natural frequencies, which are properties established by its mass and stiffness distribution.

Vibration that takes place under the excitation of external forces is called forced vibration. When the excitation is oscillatory, the system is forced to vibrate at the excitation frequency. If the excitation frequency coincides with one of the natural frequencies of the system, a condition of resonance is encountered, and dangerously large oscillations may result. Thus, the calculation of the natural frequencies of a system is of major importance in the study of vibration.

For the analysis of a vibrating system a mathematical model is required. Such a model is either discrete or continuous, the motion of which is described by coordinates. A discrete model requires a finite number of coordinates, whereas a continuous system requires an infinite number of coordinates.

The number of independent coordinates required to describe the motion of the system is called the degrees of freedom (DOF) of the system. Thus, an elastic body has an infinite number of DOF. However, such a body is often discretized to one having a finite number of DOF. In fact, a surprising number of vibration problems can be treated with sufficient accuracy by reducing a system to one having a few DOF.

II. SYSTEMS WITH ONE DEGREE OF FREEDOM

A. Free Vibration

1. Natural Frequency

The spring–mass–damper model of Fig. 1 is representative of the simplest vibration system. With its motion assumed to be restricted along a single coordinate x, the system has one DOF.

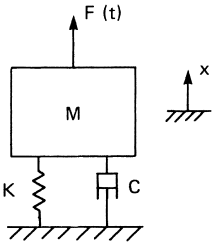


FIGURE 1 Model of a single DOFS.

The behavior of a single degree of freedom system (single DOFS) is of basic importance since coordinate transformation, to be discussed in Section III.A.3, allows systems of higher DOF to be mathematically treated in terms of equations corresponding to those of a single DOFS.

Measuring x from the equilibrium position of the mass, its differential equation of motion under excitation $F(t)$ is

$$M\ddot{x} + C\dot{x} + Kx = F(t) \tag{1}$$

where the overdots indicate time derivatives and the terms on the left side are the inertia force the damping force, and the spring force, respectively.

K is the stiffness of the spring and M is the mass. The viscous damping force $C\dot{x}$, proportional to the velocity has been assumed for mathematical convenience. In reality the damping force is not known with any degree of accuracy, and the viscous assumption enables one to find a simple solution of acceptable accuracy for small damping. Other types of damping are addressed in Section II.B.5.

For free vibration the right-hand term $F(t)$ is zero. When the natural frequency is of primary concern, the damping term is also made equal to zero, since the effect of damping on the natural frequency is generally negligible. Equation (1) then becomes

$$\ddot{x} + \omega^2 x = 0 \tag{2}$$

Here, $\omega^2 = K/M$, where ω is the circular frequency $(2\pi f)$.

This equation is that of simple harmonic moon with general solution

$$x(t) = A \sin \omega t + B \cos \omega t$$

The two arbitrary constants A and B are solved from the initial conditions $x(0)$ and $\dot{x}(0)$ for the displacement and velocity, which results in the Equation

$$x(t) = x(0) \cos \omega t + (\dot{x}(0)/\omega) \sin \omega t \tag{3}$$

In harmonic motion, a complete cycle takes place when $\omega t = 2\pi$ so that the period τ and the natural frequency of vibration become

$$\begin{aligned} &= 2\pi/\omega = 2\pi\sqrt{M/K} = \text{period} \\ &= 1/\tau = (1/2\pi)\sqrt{K/M} = \text{natural frequency} \end{aligned} \tag{4}$$

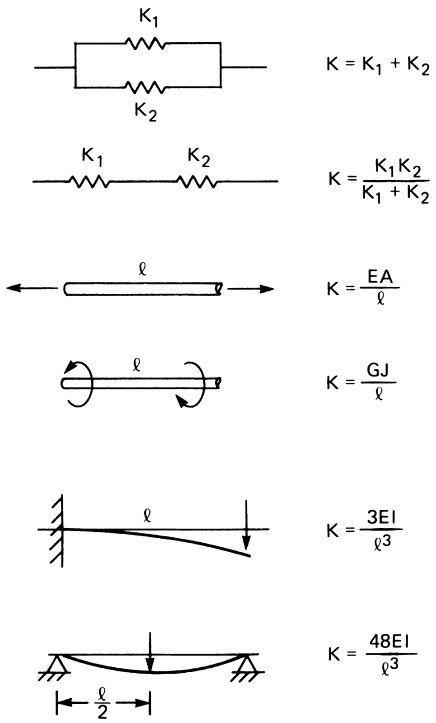


FIGURE 2 Stiffness of common elastic elements.

It is seen from Eq. (4) that the natural frequency of the single DOFS depends only on the stiffness K of the spring and the mass M.

Although the single DOFS may appear in various configurations, including rotation as well as translation, the form of the differential equation of motion is the same: that of the second-order ordinary differential equation. Stiffness K for various spring configurations is presented in Fig. 2.

2. Energy Method

It is often convenient to examine the vibration problem from energy considerations. In a conservative system the total energy must remain constant. The energies involved are kinetic energy T, due to the velocity of the mass, and the potential energy U stored in the spring.

Since the reference for U is arbitrary, it is convenient to choose it to be zero at the equilibrium position of the system. Then $U=0$ at $x=0$ and a maximum $U_{\max}$ at the extreme position $x=x_{\max}$. The kinetic energy T, on the other hand, is zero at the extreme position $x=x_{\max}$, where the velocity of the mass is zero, and a maximum $T_{\max}$ as it passes through the equilibrium position $x=0$. Thus, the principle of conservation of energy requires that

$$T_{\max} = U_{\max} \tag{5}$$

Assuming sinusoidal motion $x = A\sin\omega t$,

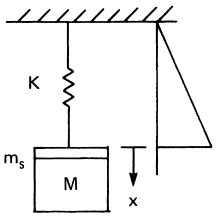


FIGURE 3 Equivalent mass of a spring.

$$\begin{aligned} T_{\max} &= \left(\tfrac{1}{2}M\dot{x}^2\right)_{\max} = \tfrac{1}{2}M\omega^2A^2 \\ U_{\max} &= \left(\tfrac{1}{2}Kx^2\right)_{\max} = \tfrac{1}{2}KA^2 \end{aligned} \tag{6}$$

Equating the two, the natural frequency is obtained as

$$\omega^2 = K/M$$

If the differential equation of motion is also desired, it can be obtained from $(d/dt)(T+U)=0$.

The simple relationship $\omega=\sqrt{K/M}$ of the single-DOF lumped-mass system can sometimes be extended to account for the mass in the elastic element for a more accurate value of the natural frequency. This is accomplished by assuming some reasonable deflection distribution of the mass in the elastic element and calculating its kinetic energy to establish an equivalent mass to add to the lumped mass M.

EXAMPLE. If the deflection of the spring in Fig. 3 is assumed to vary linearly from zero at the fixed end to x at the point of attachment to the mass M, the kinetic energy of the spring can be calculated to be $T_s=\frac{1}{2}(m_s/3)\omega^2A^2$. where m_s is the total mass of the spring. The equivalent mass is one-third the mass of the spring.

The equation for the natural frequency of the spring–mass system including the mass of the spring then becomes

$$\omega = \sqrt{K/\left(M+\tfrac{1}{3}m_s\right)}$$

Similarly, the equivalent mass of a simply supported beam of Fig. 4 can be shown to be $0.4857m_b$. For its calculation the statical deflection of the beam has been assumed, and its kinetic energy was expressed in terms of the deflection at midspan, where the beam stiffness is $K=48EI/l^3$. The equation for the natural frequency including the mass m_b of the beam is then

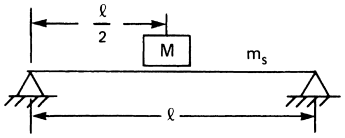


FIGURE 4 Equivalent mass of a simply supported beam.

$$\omega = \sqrt{\frac{48EI/l^3}{M + 0.4857m_b}}$$

3. Time Response with Damping

The influence of damping on the free vibration can be studied by examining the homogeneous equation

$$M\ddot{x} + C\dot{x} + Kx = 0 \tag{7}$$

The usual treatment of this equation is to assume a solution of the form $x = e^{st}$. Its substitution into the differential equation results in the characteristic equation

$$Ms^2 + Cs + K = 0 \tag{8}$$

The two roots of this equation are

$$s_{1.2} = -C/2M \pm \sqrt{(C/2M)^2 - K/M} \tag{9}$$

and the behavior of the system is dependent on the numerical value of the radical, which can be zero, positive or imaginary. Only if the radical is imaginary will the system be oscillatory.

When the radical is zero,

$$(C/2M)^2 = K/M$$

This value of C is called critical damping,

$$C_c = 2\sqrt{KM} = 2M\omega \tag{10}$$

which represents the limiting case between oscillatory and nonoscillatory motion. It is convenient, then, to consider all cases in terms of the critical damping C_c by introducing a nondimensional term $\zeta = C/C_c$, which is called the damping ratio or damping factor. The two roots of s can then be written

$$s_{1.2} = (-\zeta \pm i\sqrt{1 - \zeta^2})\omega \tag{11}$$

and the three cases of damping previously mentioned depend on whether ζ is greater than, less than, or equal to unity.

Of greatest interest is the oscillatory case in which ζ is less than 1. One form of the general solution for $\zeta < 1$ is

$$x(t) = Ae^{-\zeta\omega t}\sin(\sqrt{1 - \zeta^2}\omega t + \phi) \tag{12}$$

where A and ϕ are arbitrary constants. It is evident from this equation that the frequency of the damped oscillation is slightly lowered by damping and is equal to

$$\omega_d = \omega\sqrt{1 - \zeta^2}$$

Figure 5 shows a typical plot of a damped oscillation. The decay of oscillation shown here leads to another convenient measure of damping. Defining the natural logarithm of the ratio of any two successive amplitudes as the logarithmic decrement δ, it is easily shown from Eq. (12) that

$$\delta = \ln(x_i/x_{i+1}) = \zeta\omega\tau_d$$

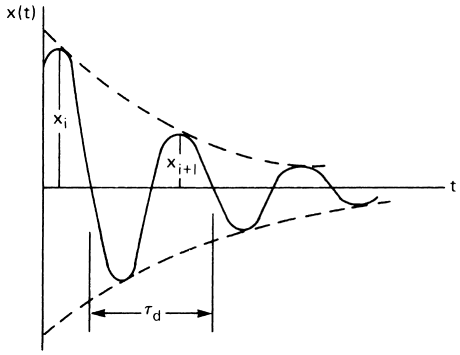


FIGURE 5 Damped free vibration.

Substituting for the damped period $\tau_d = 2\pi\omega\sqrt{1 - \zeta^2}$ the expression for the logarithmic decrement becomes

$$\delta = 2\pi\zeta/\sqrt{1 - \zeta^2} \cong 2\pi\zeta \tag{13}$$

B. Forced Vibration

1. Harmonic Excitation

When a system is subjected to forced harmonic excitation, it vibrates at the same frequency as that of the excitation. If the excitation frequency coincides with the natural frequency of the system, large amplitudes may result, and dampers and absorbers are often used to prevent dangerous conditions.

With the system of Fig. 1 excited by a harmonic force $F(t) = F_0 \sin\omega t$, the differential equation of motion is

$$M\ddot{x} + C\dot{x} + Kx = F_0 \sin\omega t \tag{14}$$

The solution to this equation can be written in the form

$$x(t) = X\sin(\omega t + \phi) \tag{15}$$

where the amplitude X and the phase ϕ of the displacement with respect to the force are given as

$$X = F_0/\sqrt{(K - M\omega^2)^2 + (C\omega)^2}$$
$$\phi = \tan^{-1}\frac{C\omega}{K - M\omega^2} \tag{16}$$

These results are shown graphically in Fig. 6. It is seen that the amplitude at resonance $\omega = \omega_1$ increases with decreasing damping and becomes infinite when the damping ratio $\zeta = 0$.

2. Rotating Unbalance

Vibration excitation may be the result of unbalance in rotating machinery.

As seen in Fig. 7, the unbalance is represented by an eccentric mass m, with eccentricity e that is rotating with angular velocity ω. The exciting force will then be

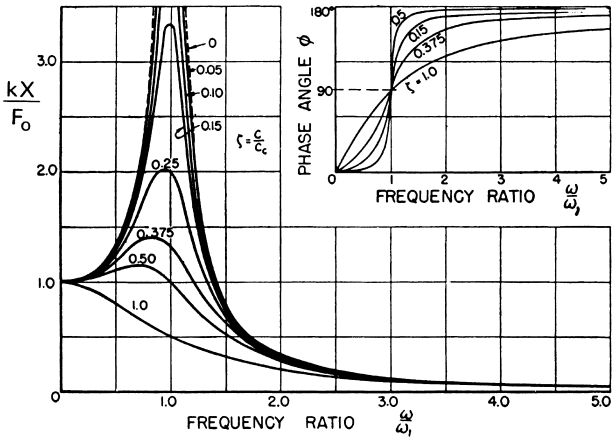


FIGURE 6 Response due to force excitation.

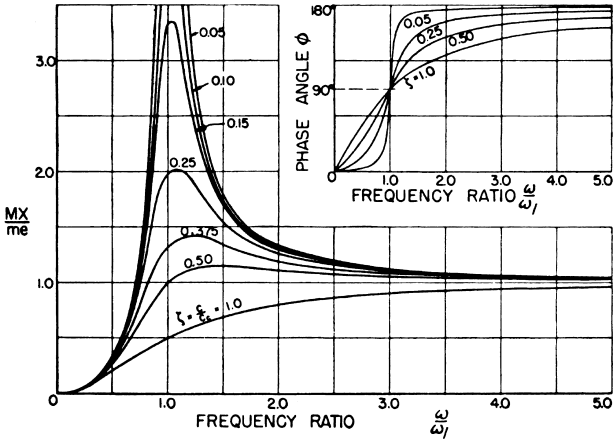


FIGURE 8 Response due to rotating unbalance.

$F = me\omega^2 \sin \omega t$, so that one need only replace F_0 in the previous section by $me\omega^2$, or

$$X = me\omega^2 / \sqrt{(K - M\omega^2)^2 + (C\omega)^2} \qquad (17)$$

The phase angle is not altered, and the response can be plotted as in Fig. 8.

3. Support Excitation

In the case where the system is excited by the motion of the support point, as shown in Fig. 9 the equation of motion becomes

$$M\ddot{x} = -K(x - y) - C(\dot{x} - \dot{y})$$

Making the substitution $z = (x - y)$, the equation can be rewritten as

$$M\dot{z} + C\dot{z} + Kz = -M\ddot{y} = M\omega^2 Y \sin \omega t \qquad (18)$$

where $y = Y \sin \omega t$ has been assumed for the motion of the base. This equation is similar in form to that of the force excitation, where z replaces x, and $M\omega^2 Y$ replaces F_0. Thus, a similar solution is expected for the relative displacement z. When Z is expressed in terms of X and Y, the result is a somewhat different equation.

$$\left|\frac{X}{Y}\right| = \sqrt{\frac{K^2 + (\omega C)^2}{(K - \omega^2 M)^2 + (\omega C)^2}} \qquad (19)$$

which is plotted in Fig. 10. It should be noted that the amplitude curves for different damping all have $|\frac{X}{Y}| = 1$ at $|\frac{\omega}{\omega_1}| = \sqrt{2}$.

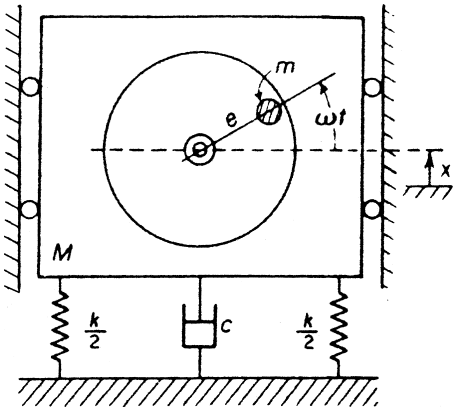


FIGURE 7 Harmonic disturbing force resulting from rotating unbalance.

4. Vibration Isolation

The results of the last section enable one to understand the basis of vibration isolation. Vibration isolation attempts either to protect a delicate object from excessive vibration transmitted to it from its supporting structure or to prevent vibratory forces generated by machines from being transmitted to their surroundings. The basic problem is the same for these two objectives—that of reducing the transmitted forces.

Figure 10 shows that the ratio $|X/Y|$ for the motion transmitted from the vibrating floor to the supported mass is less than 1.0 when the frequency ratio ω/ω_1 is greater than $\sqrt{2}$. This means that the natural frequency $\omega_1 = \sqrt{K/M}$ of the suspended system must be low in

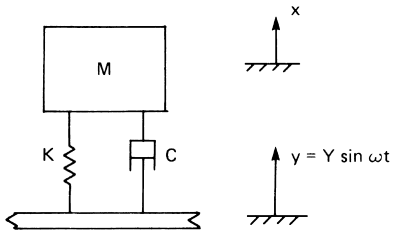


FIGURE 9 System under support motion.

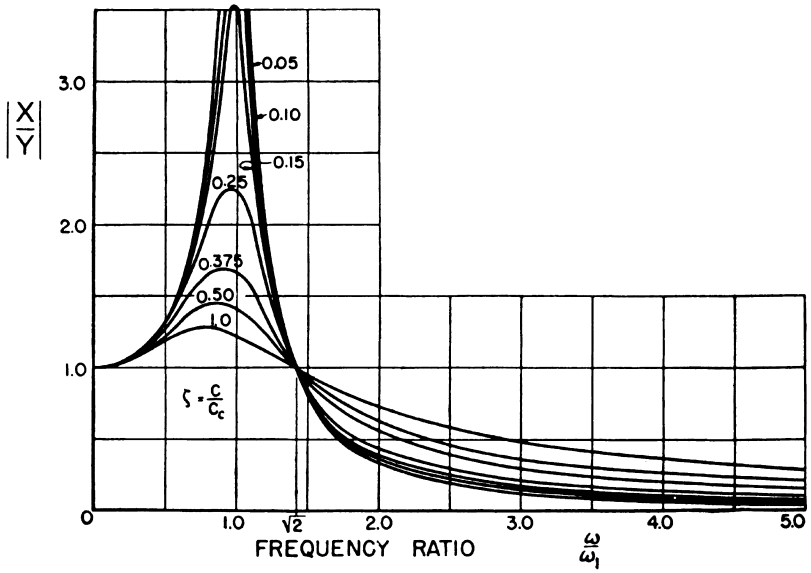


FIGURE 10 Response due to support excitation.

comparison with the disturbing frequency ω. This can be accomplished by a soft spring (small K).

The second problem, that of reducing the force transmitted by a machine to the floor, has the same requirement. The force to be isolated is transmitted by the spring–damper system, which has the value

$$F_T = \sqrt{(KX)^2 + (C\omega X)^2} = X\sqrt{K^2 + (C\omega)^2}$$

Substituting for the displacement X produced by the exciting force F, which is

$$X = F/\sqrt{(K - M\omega^2)^2 + (C\omega)^2}$$

the ratio of the transmitted force F_T to the exciting force F is

$$F_T/F = \sqrt{K^2 + (\omega C)^2}/\sqrt{(K - M\omega^2)^2 + (C\omega)^2} \quad (20)$$

which is identical to the equation for $|X/Y|$ plotted as Fig. 10.

Another solution to either problem is to mount the system to be isolated on a heavy block supported on a cushion of soft material like spongy rubber or felt. In this way M is increased to lower the natural frequency and increase the frequency ratio ω/ω_1.

Since in the general problem the mass to be isolated has six DOF (three translation and three rotation), the designer of the isolator system must use intuition and ingenuity. The results of the single-DOF analysis should however, serve as a useful guide.

5. Equivalent Damping

Damping is present in all oscillatory systems. The decay of amplitude in free vibration is due to the loss of energy by damping. In forced steady-state vibration the loss of energy is balanced by the energy supplied by the excitation force.

There are many different kinds of damping forces, from internal molecular friction to sliding friction and fluid resistance. Their exact mathematical description is difficult, and a practical approach is to utilize the concept of equivalent viscous damping based on equal energy dissipation.

We need for this, w_d the energy dissipated per cycle by viscous damping under harmonic oscillation $x = X \sin(\omega t - \phi)$. Its substitution into the work equation results in

$$W_d = \oint (C\dot{x})\,dx = \oint C\dot{x}^2\,dt = \pi C\omega X^2$$

For the equivalent viscous damping, we then write

$$\pi C_{eq}\omega X^2 = W_d \quad (21)$$

where W_d is now the energy dissipated by any damping force. Of course, a simple relationship for W_d may not be available, but C_{eq} found in this manner enables one to use all the elementary equations of forced vibration of the previous sections.

Mentioned here is just one form of damping, that of solid damping encountered by many structural materials such as steel and aluminum and often referred to as structural or hysteresis damping. For these materials the energy dissipated per cycle is independent of the frequency over a wide frequency range and proportional to the square of the amplitude of vibration.

$$W_d = \alpha X^2 \quad (22)$$

By comparsion of W_d in the two equations, the equivalent viscous damping coefficient is $C_{eq} = \alpha/\pi\omega$, and the damping force becomes

$$F_d = (\alpha/\pi)X\cos(\omega t - \phi)$$

Thus, for structural damping, the amplitude and phase under steady-state harmonic force becomes

$$X = F/\sqrt{(K - M\omega^2)^2 + (\alpha/\pi)^2}$$

$$\phi = \tan^{-1}\frac{\alpha/\pi}{K - M\omega^2}$$

By letting $\gamma = \alpha/\pi K$, the differential equation of motion can be written

$$M\ddot{x} + K(1 + i\gamma)x = Fe^{i\omega t} \tag{23}$$

The quantity $K(1 + i\gamma)$ is called the complex stiffness and γ the structural damping factor. The relation between γ and the viscous damping factor ζ is easily found by comparing the response at resonance:

$X = F/2\zeta K$, amplitude at resonance (viscous)

$X = F/\gamma K$, amplitude at resonance (structural)

Thus, it can be concluded that the structural damping factor γ is equal to twice the viscous damping factor ζ on the basis of equal resonant amplitudes.

C. Transient Vibration

When a dynamic system is excited by a non-periodic force, such as a suddenly applied impact, steady-state oscillations are not produced and the resulting motion is called a transient response. The transient response of a simple spring–mass system of single DOF will be discussed in this section since its behavior is basic to the transient behavior of the more complex system.

1. Impulse Response

Impulse is the time integral of a force as expressed by the equation

$$\hat{F} = \int F(t)\,dt \tag{24}$$

when a force of very large magnitude acts for a very short time, its time integral can be finite. Such force is described as impulsive. Letting the magnitude of the impulsive force be $\hat{F}/\varepsilon$ over a time duration of ε, its limiting case $\varepsilon \to 0$ with impulse value of unity is called the unit impulse or the Dirac delta function.

A delta function at $t = \xi$ is identified by the symbol $\delta(t - \xi)$ and has the following properties:

$$\delta(t - \xi) = 0 \qquad \text{for all} \quad t \neq \xi$$

$$\int_0^\infty \delta(t - \xi)\,dt = 1.0, \qquad 0 < \xi < \infty$$

Thus, when $\delta(t - \xi)$ is multiplied by any time function $f(t)$, its product will be zero everywhere except at $t = \xi$, and its time integral will be

$$\int_0^\infty f(t)\,\delta(t - \xi)\,dt = f(\xi), \qquad 0 < \xi < \infty$$

Since impulse is equal to the change in momentum, $\hat{F}$ acting on a mass will result in a sudden change in its velocity equal to $\hat{F}/M$ without an appreciable change in its displacement. From Eq. (3) for the free vibration of an undamped spring–mass system with initial conditions $x(0)$ and $\dot{x}(0)$, it it evident that the response of the spring–mass system initially at rest and excited by an impulse $\hat{F}$ is

$$x = (\hat{F}/M\omega_1)\sin\omega_1 t \tag{25}$$

The oscillation that takes place is at the natural frequency ω_1.

Similarly, the response of a damped spring–mass system can be determined from Eq. (12) to be

$$x = \frac{\hat{F}}{M\omega_1\sqrt{1 - \zeta^2}}\,e^{-\zeta\omega_1 t}\sin\sqrt{1 - \zeta^2}\,\omega_1 t \tag{26}$$

2. Arbitrary Excitation

Letting $h(t)$ be the response to a unit impulse, the response to an impulse $\hat{F}$ becomes $x(t) = \hat{F}h(t)$. The response to an arbitrary force $f(t)$ can be established in terms of $h(t)$ by considering $f(t)$ to be a series of impulses of strength $\hat{F} = f(\xi)\,\Delta\xi$, as shown in Fig 11. Clearly, the response to the unit impulse at $t = \xi$ is

$$f(\xi)\,\Delta\xi\,h(t - \xi)$$

where $(t - \xi)$ is the elapsed time after the impulse. For a linear system the principle of super-position holds, and the total response at time t is found by summing all such contributions as

$$x(t) = \int_0^t f(\xi)h(t - \xi)\,d\xi \tag{27}$$

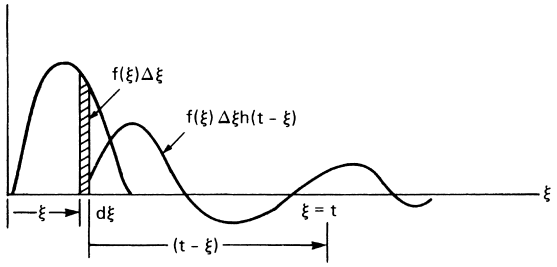


FIGURE 11 Response to arbitrary excitation.

This integral is known as the *convolution integral*. Since $f(\xi) = 0$ for times greater than the pulse duration t_p, the upper limit of the integral remains constant at t_p for $t > t_p$. Another form of this integral can be written by substituting $\tau = t - \xi$, which leads to

$$x(t) = \int_0^t f(t-\tau)h(\tau)\,d\tau \tag{28}$$

EXAMPLE. Determine the response of an undamped single DOFS to a step function of magnitude F_0. The response to a unit impulse is $h(t) = (1/M\omega_1)\sin\omega_1 t$. Substituting into the convolution integral, the result is

$$\begin{aligned} x(t) &= \frac{F_0}{M\omega_1}\int_0^t \sin\omega_1(t-\xi)\,d\xi \\ &= (F_0/K)(1-\cos\omega_1 t) \end{aligned}$$

If the response of a damped system is desired, this procedure is repeated with

$$h(t) = \frac{e^{-\zeta\omega_1 t}}{m\omega_1\sqrt{1-\zeta^2}}\sin\sqrt{1-\zeta^2}\omega_1 t \tag{29}$$

III. SYSTEMS WITH MULTIPLE DEGREES OF FREEDOM

A multi-DOFS requires more than one coordinate to describe its motion. Such systems differ from the single DOFS in that an n DOFS has n natural frequencies, and for each of the natural frequencies there corresponds a natural state of vibration with a displacement configuration known as normal mode. Mathematical terms for these quantities are known as eigenvalues and eigenvectors. They are established from the n simultaneous differential equations of motion and possess certain dynamic properties associated with the system.

The normal mode vibrations are free vibrations that depend only on the mass and stiffness distribution of the system. They are of importance not only in establishing the spectrum of resonance but also in the fact that forced vibration can be analyzed in terms of normal modes.

For an n DOFS there will be a set of n simultaneous differential equations to solve. To carry out this task efficiently, matrix methods are employed. They provide a compact notation and an organized method for the solution of linear simultaneous equations.

A. Two Degrees of Freedom

Since the two DOFS is a special case of the multi-DOFS, all of the concepts of the multi-DOFS apply to the two DOFS. Its discussion at this point is warranted in that simple analytic solutions with numerical examples are easily obtained for the two DOFS, whereas this is not the case for the larger systems. Computers are manditory for higher order multi-DOFS.

A two DOFS has two natural frequencies at which the motion displays two distinct modes of oscillation called normal modes. These normal modes can be produced by proper initial conditions. For a more general initial condition, the free vibration produced will be the superposition of the two normal modes. As in the single DOFS, forced harmonic motion will take place at the excitation frequency, and the amplitude will increase to a maximum at the two natural frequencies.

1. Natural Frequencies and Mode Shapes

The system shown in Fig. 12 requires two coordinates to describe its motion and hence it has two DOF. Applying Newton's laws of motion, we can write the following two equations:

$$\begin{aligned} m_1\ddot{x}_1 &= -k_1x_1 + k_2(x_2 - x_1) \\ m_2\ddot{x}_2 &= -k_2(x_2 - x_1) - k_3x_2 \end{aligned}$$

These can be expressed in matrix notation as

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}\begin{Bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{Bmatrix} + \begin{bmatrix} (k_1+k_2) & -k_2 \\ -k_2 & (k_2+k_3) \end{bmatrix}\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{30}$$

For the normal mode vibration, each mass undergoes harmonic motion of the same frequency ω. To find the normal mode frequencies substitute $x_j = A_j e^{i\omega t}$ for $j = 1, 2$ into Eq. (30) and Eq. (30) becomes

$$\begin{bmatrix} (k_1+k_2-\omega^2 m_1) & -k_2 \\ -k_2 & (k_2+k_3-\omega^2 m_2) \end{bmatrix}\begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} \tag{31}$$

The natural frequencies ω_1 and ω_2 are found from the characteristic equation given by the determinant

$$\begin{vmatrix} (k_1+k_2-\omega^2 m_1) & -k_2 \\ -k_2 & (k_2+k_3-\omega^2 m_1) \end{vmatrix} = 0 \tag{32}$$

and the normal modes are found by solving for the ratio

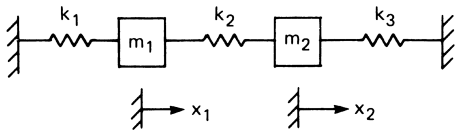


FIGURE 12 System with two DOF.

x_1/x_2 in any one of the equations of motion and presented as a column matrix $\{{}^{x_1}_{x_2}\}$.

EXAMPLE. Illustrating by numbers, assume the stiffness and the mass to be equal to k and m. We then have

$$\begin{bmatrix}(2-\omega^2 m/k) & -1\\ -1 & (2-\omega^2 m/k)\end{bmatrix}\begin{Bmatrix}x_1\\ x_2\end{Bmatrix}=\begin{Bmatrix}0\\ 0\end{Bmatrix}$$

Letting $\lambda=\omega^2 m/k$, the characteristic equation is

$$\lambda^2-4\lambda+3=0$$

which is satisfied by $\lambda=1$ and $\lambda=3$. The two natural frequencies are then obtained from

$$\omega_1^2=k/m$$
$$\omega_2^2=3k/m$$

Substituting these back into the equations of motion, the normal modes are found to be

$$\text{Mode 1:}\quad \begin{Bmatrix}x_1\\ x_2\end{Bmatrix}_1=\begin{Bmatrix}1\\ 1\end{Bmatrix}$$

$$\text{Mode 2:}\quad \begin{Bmatrix}x_1\\ x_2\end{Bmatrix}_2=\begin{Bmatrix}1\\ -1\end{Bmatrix}$$

In a liner system, the principle of superposition holds, and the sum of the normal modes will also be a solution. Thus, any free vibration for this system can be written

$$\begin{Bmatrix}x_1\\ x_2\end{Bmatrix}=A_1\begin{Bmatrix}1\\ 1\end{Bmatrix}\sin(\omega_1 t-\varphi_1)+A_2\begin{Bmatrix}1\\ -1\end{Bmatrix}\sin(\omega_2 t-\varphi_2)$$

where the arbitrary constants A_i and φ_i are determined by the initial conditions.

2. Choice of Coordinates

In general, the equations of motion

$$[M]\{\ddot{q}\}+[K]\{q\}=\{0\} \tag{33}$$

are coupled. If both the M and K matrices in Eq. (33) are diagonal, the equations of motion are decoupled, and each equation can be solved independently, as in the single-DOF case. Thus, coupling of coordinates arises from the off-diagonal terms.

When off-diagonal terms appear in the mass matrix, the system is said to have dynamic coupling. This is equivalent to having cross-products of coordinates in the equation for the kinetic energy.

If off-diagonal terms appear in the stiffness matrix, the system is said to have static coupling. A system with static coupling will have cross-products of coordinates in the potential energy equation.

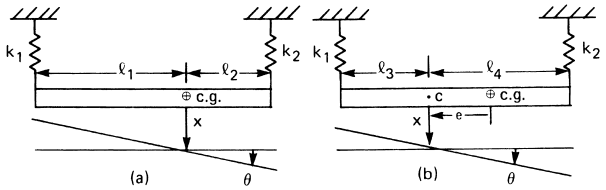


FIGURE 13 Choice of coordinates (c.g., center of gravity).

EXAMPLE. In this example coordinates are chosen in two different ways, leading to static coupling and dynamic coupling.

In Fig. 13a the displacement x is chosen at the center of gravity of the bar. Its equation of motion,

$$\begin{bmatrix}m & 0\\ 0 & J_{\text{c.g.}}\end{bmatrix}\begin{Bmatrix}\ddot{x}\\ \ddot{\theta}\end{Bmatrix}+\begin{bmatrix}(k_1+k_2) & (k_2l_2-k_1l_1)\\ (k_2l_2-k_1l_1) & (k_1l_1^2+k_2l_2^2)\end{bmatrix}\begin{Bmatrix}x\\ \theta\end{Bmatrix}=\begin{Bmatrix}0\\ 0\end{Bmatrix}$$

where $J_{\text{c.g.}}$ is the mass moment of inertia, shows static coupling.

In Fig. 13b a point c along the bar is chosen where a force applied normal to the bar produces pure translation; that is, $k_1l_3=k_2l_4$. Its equation of motion,

$$\begin{bmatrix}m & me\\ me & J_e\end{bmatrix}\begin{Bmatrix}\ddot{x}\\ \ddot{\theta}\end{Bmatrix}+\begin{bmatrix}(k_1+k_2) & 0\\ 0 & (k_1l_3^2+k_2l_4^2)\end{bmatrix}\begin{Bmatrix}x\\ \theta\end{Bmatrix}=\begin{Bmatrix}0\\ 0\end{Bmatrix}$$

shows dynamic coupling.

If the coordinate x is measured at the end of the bar, both the mass and stiffness matrices will be full and dynamic, and static coupling will appear simultaneously.

The choice of coordinates is arbitrary and does not affect the nature of the vibration. Regardless of the coordinates chosen, the two natural frequencies and their normal modes remain unchanged.

3. Normal Coordinates

It has been demonstrated that the elements of the mass and stiffness matrices depend on the choice of coordinates. It can be shown that there is a set of coordinates, called *principle* or *normal coordinates*, that will diagonalize the M and K matrices and thereby decouple the equations of motion.

EXAMPLE. For the system of Fig. 12 with $k_i=k$ and $m_i=m$, the equation of motion

$$m\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{Bmatrix}\ddot{x}_1\\ \ddot{x}_2\end{Bmatrix}+k\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{Bmatrix}x_1\\ x_2\end{Bmatrix}=\begin{Bmatrix}0\\ 0\end{Bmatrix}$$

shows static coupling. When written out, these are

$$m\ddot{x}_1 + 2kx_1 - kx_2 = 0$$
$$m\ddot{x}_2 - kx_1 + 2kx_2 = 0$$

Adding and subtracting these equations, we obtain a new set of equations:

$$m(\ddot{x}_1 + \ddot{x}_2) + 2k(x_1 + x_2) - k(x_1 + x_2) = 0$$
$$m(\ddot{x}_1 - \ddot{x}_2) + 2k(x_1 - x_2) + k(x_1 - x_2) = 0$$

Thus, if we let

$$p_1 = x_1 + x_2$$
$$p_2 = x_1 - x_2$$

the above equations become

$$m\ddot{p}_1 + kp_1 = 0$$
$$m\ddot{p}_2 + 3kp_2 = 0$$

or in matrix notation

$$m\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{Bmatrix}\ddot{p}_1\\ \ddot{p}_2\end{Bmatrix} + k\begin{bmatrix}1 & 0\\ 0 & 3\end{bmatrix}\begin{Bmatrix}p_1\\ p_2\end{Bmatrix} = \begin{Bmatrix}0\\ 0\end{Bmatrix}$$

The equations in the normal coordinates p are now decoupled, with each equation corresponding to one of single DOF.

The transformation between the x and the p coordinates in matrix form is

$$\begin{Bmatrix}p_1\\ p_2\end{Bmatrix} = \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{Bmatrix}x_1\\ x_2\end{Bmatrix}$$

and its inverse is

$$\begin{Bmatrix}x_1\\ x_2\end{Bmatrix} = \frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{Bmatrix}p_1\\ p_2\end{Bmatrix}$$

Note that each column of the transformation matrix corresponds to the normal modes of the system.

For a more complex set of equations, this technique for finding a set of normal coordinates would not be practical. In Section III.B.4, after a discussion of the orthogonality of normal modes, a more general method or normalizing the equations of motion will be presented.

4. Forced Harmonic Motion

The matrix equation for the two DOFS excited by a harmonic motion acting on mass m_1 is

$$\begin{bmatrix}m_1 & 0\\ 0 & m_2\end{bmatrix}\begin{Bmatrix}\ddot{x}_1\\ \ddot{x}_2\end{Bmatrix} + \begin{bmatrix}k_{11} & k_{12}\\ k_{21} & k_{22}\end{bmatrix}\begin{Bmatrix}x_1\\ x_2\end{Bmatrix} = \begin{Bmatrix}F_1\\ 0\end{Bmatrix}\sin\omega t \tag{34}$$

Since in forced vibration the system responds at the same frequency as that of the excitation, we can assume the solution in the form

$$\begin{Bmatrix}x_1\\ x_2\end{Bmatrix} = \begin{Bmatrix}X_1\\ X_2\end{Bmatrix}\sin\omega t$$

Substituting this equation into the equation of motion, we obtain

$$\begin{bmatrix}(k_{11} - m_1\omega^2) & k_{12}\\ k_{21} & (k_{22} - m_2\omega^2)\end{bmatrix}\begin{Bmatrix}X_1\\ X_2\end{Bmatrix} = \begin{Bmatrix}F_1\\ 0\end{Bmatrix} \tag{35}$$

which can be abbreviated as

$$[Z(\omega)]\begin{Bmatrix}X_1\\ X_2\end{Bmatrix} = \begin{Bmatrix}F_1\\ 0\end{Bmatrix}$$

Premultiplying by the inverse and noting that $[Z(\omega)]^{-1} = \text{adj}[Z(\omega)]/|Z(a)|$, we obtain

$$\begin{Bmatrix}X_1\\ X_2\end{Bmatrix} = [Z(\omega)]^{-1}\begin{Bmatrix}F_1\\ 0\end{Bmatrix} = \frac{\text{adj}[Z(\omega)]}{|Z(\omega)|}\begin{Bmatrix}F_1\\ 0\end{Bmatrix} \tag{36}$$

Here, adj[ ] denotes adjoint matrix.

To express this equation in another form, we note that $|Z(\omega)| = 0$ is the characteristic equation yielding the roots ω_1 and ω_2. Thus, it is possible to rewrite $|Z(\omega)|$ as

$$|Z(\omega)| = m_1m_2(\omega_1^2 - \omega^2)(\omega_2^2 - \omega^2)$$

The adj$[Z(\omega)]$ is also

$$\begin{bmatrix}(k_{22} - m_2\omega^2) & -k_{12}\\ -k_{21} & (k_{11} - m_1\omega^2)\end{bmatrix}$$

Thus, the equations for X_1 and X_2 become

$$\begin{Bmatrix}X_1\\ X_2\end{Bmatrix} = \frac{\begin{bmatrix}(k_{22} - m_2\omega^2) & -k_{12}\\ -k_{21} & (k_{11} - m_1\omega^2)\end{bmatrix}\begin{Bmatrix}F_1\\ 0\end{Bmatrix}}{m_1m_2(\omega_1^2 - \omega^2)(\omega_2^2 - \omega^2)} \tag{37}$$

EXAMPLE. The equation of motion for the system shown in Fig. 14 is

$$\begin{bmatrix}m & 0\\ 0 & m\end{bmatrix}\begin{Bmatrix}\ddot{x}_1\\ \ddot{x}_2\end{Bmatrix} + \begin{bmatrix}2k & -k\\ -k & 2k\end{bmatrix}\begin{Bmatrix}x_1\\ x_2\end{Bmatrix} = \begin{Bmatrix}F_1\\ 0\end{Bmatrix}\sin\omega t$$

Thus, $k_{11} = k_{22} = 2k$; $k_{12} = k_{21} = -k$; $\omega_1^2 = k/m$ and $\omega_2^2 = 3k/m$. Then X_1 and X_2 become

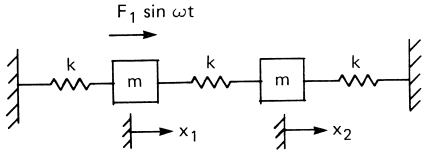


FIGURE 14 Forced vibration of two DOFS.

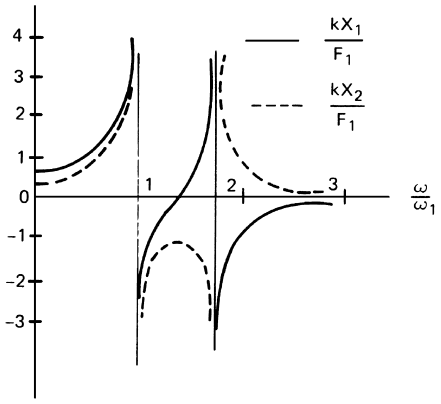


FIGURE 15 Forced vibration of system of Fig. 14.

$$X_1 = \frac{(2k - m\omega^2)F_1}{m^2(\omega_1^2 - \omega^2)(\omega_2^2 - \omega^2)}$$

$$X_2 = \frac{kF_1}{m^2(\omega_1^2 - \omega^2)(\omega_2^2 - \omega^2)}$$

These equations plotted in nondimensional form are shown in Fig. 15.

B. Higher Degrees of Freedom

1. General Equations of Motion

We now introduce the equations of motion in a more general notation:

$$\begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \begin{Bmatrix} \ddot{q}_1 \\ \ddot{q}_2 \\ \ddot{q}_3 \end{Bmatrix} + \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{Bmatrix} \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \end{Bmatrix} + \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix} \begin{Bmatrix} q_1 \\ q_2 \\ q_3 \end{Bmatrix} = \begin{Bmatrix} F_1 \\ F_2 \\ F_3 \end{Bmatrix} \tag{38}$$

In this equation, the square matrices are all symmetric about the diagonal, as was found for previous example of two DOF, that is, $m_{ij} = m_{ji}$, $c_{ij} = c_{ji}$, and $k_{ij} = k_{ji}$. This property of symmetry results from Maxwell's reciprocal theorem, which states that the work done on any linear structure by loads applied at two different points is independent of the order of loading.

Examining the terms of the stiffness matrix, the following interpretation can be made. If q_i is made equal to unity with all other q's equal to zero, Eq. (38) states that the terms of the ith column $k_{1i}, k_{2i}, k_{3i}, \ldots$ are equal to the forces $f_1, f_2, f_3, \ldots$ required at stations 1, 2, 3, ... in order to maintain this displacement configuration. Thus, the stiffness term of any column can be determined by letting the displacement corresponding to that column be unity with all other displacements equal to zero and measuring the forces required at each station.

For concise presentation of the matrix equation, the form

$$[M]\{\ddot{q}\} + [C]\{\dot{q}\} + [K]\{q\} = \{F\}$$

is generally used, and when there is no ambiguity, brackets and braces are often dispensed with, that is,

$$M\ddot{q} + C\dot{q} + Kq = F \tag{39}$$

Similarity with the equation for the single DOF is strikingly evident.

2. Eigenvalues and Eigenvectors

For the calculation of the natural frequencies and mode shapes, the damping terms are deleted along with the forcing terms, the equation taking the form

$$M\ddot{q} + Kq = 0$$

Since the normal modes execute harmonic motion, $\ddot{q} = -\omega^2 q$. By letting $\lambda = \omega^2$, the equation to be solved becomes

$$[K - \lambda M]\{q\} = \{0\} \tag{40}$$

The characteristic equation is the determinant of the equation equated to zero, or

$$|K - \lambda M| = 0 \tag{41}$$

and the roots λ_i of this equation are the eigen-values of the system. The normal modes (eigen-vectors) are then found by substituting the λ's back into the matrix equation.

3. Orthogonality of Modes

The normal modes of the system can be shown to be orthogonal with respect to the mass and the stiffness matrices. Letting the normal modes for the ith mode be represented as $\{u\}_i$, the equation for the ith mode can be written

$$\lambda_i[M]\{u\}_i = [K]\{u\}_i \tag{42}$$

Premultiplying the ith equation by the transpose of $\{u\}_j$, we obtain

$$\lambda_i\{u\}_j^{\mathrm{T}}[M]\{u\}_i = \{u\}_j^{\mathrm{T}}[K]\{u\}_i$$

If we start with the equation for the jth mode and repeat the above operation. we obtain a similar equation with the i and j interchanged. Now by subtracting the two equations and noting that for symmetric matrices

$$\{u\}_j^{\mathrm{T}}[\ ]\{u\}_i = \{u\}_i^{\mathrm{T}}[\ ]\{u\}_j$$

we obtain

$$(\lambda_i - \lambda_j)\{u\}_i^{\mathrm{T}}[M]\{u\}_j = 0$$

Since λ_i differs from λ_j, the above equation requires that

$$\{u\}_i^{\mathrm{T}}[M]\{u\}_j = 0 \quad \text{for} \quad i \neq j \tag{43}$$

Examining the original equation for the ith mode multiplied by the transpose of the jth mode, we also conclude that

$$\{u\}_i^{\mathrm{T}}[K]\{u\}_j = 0 \qquad \text{for} \quad i \neq j \tag{44}$$

These equations define the orthogonal property of normal modes.

Finally if $i = j$, $(\lambda_i - \lambda_j) = 0$ and

$$\begin{aligned} \{u\}_j^{\mathrm{T}}[M]\{u\}_j &= M_j \\ \{u\}_j^{\mathrm{T}}[K]\{u\}_j &= K_j \end{aligned} \tag{45}$$

where M_j and K_j can be any finite quantity. These are called the generalized mass and the generalized stiffness.

EXAMPLE. The natural frequencies and normal modes for the system of Fig. 12 were found as

$$\omega_1^2 = \frac{k}{m}, \quad \{u\}_1 = \begin{Bmatrix} 1 \\ 1 \end{Bmatrix}; \quad \omega_2^2 = \frac{3k}{m}, \quad \{u\}_2 = \begin{Bmatrix} 1 \\ -1 \end{Bmatrix}$$

The M and K matrices for the system are

$$M = m\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \qquad K = k\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

The orthogonality relations then have the following values:

$$(1 \quad -1)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} = 0$$

$$(1 \quad -1)\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} = 0$$

The values of the generalized mass and the generalized stiffness are

$$M_1 = (1 \quad 1)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} = 2$$

$$M_2 = (1 \quad -1)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{Bmatrix} 1 \\ -1 \end{Bmatrix} = 2$$

$$K_1 = (1 \quad 1)\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{Bmatrix} 1 \\ 1 \end{Bmatrix} = 2$$

$$K_2 = (1 \quad -1)\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{Bmatrix} 1 \\ -1 \end{Bmatrix} = 6$$

4. Modal Matrix

There are several computer programs available to solve the undamped homogeneous equation,

$$M\ddot{q} + Kq = 0$$

for its natural frequencies and normal modes. By using these results and forming a model matrix P, the general equation for the forced vibration can be decoupled and solved as a system of equations corresponding to those of the single DOFS.

The modal matrix is composed of columns of normal modes $\{u\}_i$ as follows:

$$P = [\{u\}_1\{u\}_2\{u\}_3] \tag{46}$$

If each of the modal columns is divided by the generalized mass M_i of the mode, a weighted modal matrix $\tilde{P}$ is formed.

If we use the coordinate transformation $q = \tilde{P}_p$ and premultiply by the transpose $\tilde{P}^{\mathrm{T}}$, the general equation

$$M\ddot{q} = C\dot{q} + Kq = F$$

for the forced vibration becomes

$$\tilde{P}^{\mathrm{T}}M\tilde{P}\ddot{p} + \tilde{P}^{\mathrm{T}}C\tilde{P}\dot{q} + \tilde{P}^{\mathrm{T}}K\tilde{P}q = \tilde{P}^{\mathrm{T}}F$$

Due to the orthogonality of the normal modes, the mass and stiffness matrices are diagonalized:

$$\tilde{P}^{\mathrm{T}}M\tilde{P} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} = I \quad \text{(unit matrix)} \tag{47}$$

$$\tilde{P}^{\mathrm{T}}k\tilde{P} = \begin{bmatrix} \omega_1^2 & & \\ & \omega_2^2 & \\ & & \omega_3^2 \end{bmatrix} = \Lambda \quad \begin{matrix} \text{(diagonal} \\ \text{matrix of} \\ \text{natural} \\ \text{frequencies)} \end{matrix} \tag{48}$$

The damping matrix, however, is in general not diagonalized, and the system of equations remains coupled by damping.

If the damping matrix is proportional to the mass or stiffness matrix, it is evident that the matrix $\tilde{P}^{\mathrm{T}}C\tilde{P}$ is also diagonalized. The C matrix is then called proportional damping, and each of the decoupled equations will be of the form

$$\ddot{p}_i + 2\zeta_i\omega_i\dot{p}_i + \omega_i^2 p_i = f_i(t) \tag{49}$$

corresponding to that of the single DOFS.

When C can be expressed in the form

$$C = \alpha M + \beta K \tag{50}$$

known as Rayleigh damping, where α and β are constants, the forced vibration equation can again be decoupled by the transformation $q = \tilde{P}p$. In this case $\tilde{P}^{\mathrm{T}}C\tilde{P}$ becomes

$$\tilde{P}^{\mathrm{T}}C\tilde{P} = \alpha\tilde{P}^{\mathrm{T}}M\tilde{P} + \beta\tilde{P}^{\mathrm{T}}K\tilde{P} = \alpha I + \beta\Lambda$$

and each of the decoupled equations will have the form

$$\ddot{p}_i + \left(\alpha + \beta\omega_i^2\right)\dot{p}_i + \omega_i^2 p_i = f_i(t) \tag{51}$$

IV. CONTINUOUS SYSTEMS: NORMAL MODES

Systems having continuously distributed mass and elasticity lead to partial differential equations. The general solution to these equations, including transient vibration, is beyond the scope of this article. Analytic solutions for moving boundary conditions would involve the use of Laplace transformation. Numerical solutions for the more general configurations would require discretization and the aid of computers. In this section only the normal mode free vibration of a few simple bodies with the usual boundary conditions will be discussed.

A. Strings and Rods

The flexible string with uniformly distributed mass, and the uniform rod in axial or torsional vibration, lead to the same equation of motion known as the wave equation. It has the form

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \tag{52}$$

where x is the continuous coordinate along the string or rod and c the velocity of propagation of the disturbance along its length.

The propagation velocity for each case is as follows:

String under tension T:

$$c = \sqrt{T/\rho}, \qquad \rho = \text{mass per unit length}$$

Axial vibration of rod:

$$c = \sqrt{E/\rho}$$

E = Young's modulus of elasticity

ρ = mass per unit volume

Torsional vibration of rod:

$$c = \sqrt{G/\rho}, \quad G = \text{shear modulus}$$

ρ = mass per unit volume

The solution to the wave equation can be expressed as

$$u(x,t) = F_1(ct - x) + F_2(ct + x) \tag{53}$$

where F_1 and F_2 are arbitrary functions. This equation implies that a wave travels along the x axis in the forward and backward directions with the propagation velocity c.

One method of solving the partial differential equation is that of separation of variables. In this method the solution is assumed in the form

$$u(x,t) = U(x)G(t) \tag{54}$$

Its substitution into the differential equation results in the following two equations:

$$U(x) = A\sin(\omega x/c) + B\cos(\omega x/c)$$

$$G(t) = C\sin\omega t + D\cos\omega t$$

The constants A and B are solved from the boundary conditions at the two ends, whereas the constants C and D are established from initial conditions.

EXAMPLE. For a uniform rod in longitudinal vibration with one end fixed and the other end free, the boundary conditions are as follows:

Fixed end $x = 0$; displacement $U(0) = 0$ therefore

$$B = 0$$

Free end $x = l$; stress $E\left(\frac{dU}{dx}\right)_{x=l} = 0$ therefore

$$A\frac{\omega}{c}\cos\frac{\omega l}{c} = 0$$

This is satisfied when

$$\cos(\omega l/c) = 0$$

or

$$\omega l/c = \pi/2, \qquad 3\pi/2 \cdots (2n-1)\pi/2, \qquad n = 1, 2, 3, \ldots$$

B. Beams

For the lateral vibration of uniform beams, the following differential equation, known as Euler's equation, applies:

$$EI(\partial^4 y/\partial x^4) + \rho(\partial^2 y/\partial t^2) = 0 \tag{55}$$

where ρ is the mass per unit length of beam. For the normal mode vibration, $\partial^2 y/\partial t^2 = -\omega^2 y$ and Eq. (55) becomes

$$d^4 y/dx^4 - \beta^4 y = 0 \tag{56}$$

where $\beta^4 = \rho\omega^2/EI$.

Since this is a fourth-order differential equation, the general solution has four arbitrary constants:

$$y = A\cosh\beta x + B\sinh\beta x + C\cos\beta x + D\sin\beta x \tag{57}$$

These constants and the values of β must be solved from the boundary conditions. The natural frequencies can then be established from the equation

$$\omega = (\beta l)^2\sqrt{EI/\rho l^4} \tag{58}$$

Values of $(\beta l)^2$ for the first three normal modes for various boundary conditions are given in Table I.

TABLE I Numerical Values for $(\beta l)^2$

Beam configuration	First mode	Second mode	Third mode
Simply supported	9.87	39.5	88.9
Cantilever	3.52	22.0	61.7
Free–free	22.4	61.7	121.0
Clamped–clamped	22.4	61.7	121.0
Clamped–hinged	15.4	50.0	104.0

V. LAGRANGE'S EQUATION: GENERALIZED COORDINATES

For complex problems, the vector method based on Newton's laws becomes unwieldy, and the scalar method based on energy should be considered. The usefulness of the energy method has already been demonstrated for the single DOFS. However, the statement for the total energy, as used for the single DOFS, provides only one equation, which is insufficient for multi-DOFS. This limitation was overcome by Joseph L. C. Lagrange (1736–1813).

A. Equations of Motion

Lagrange developed a general treatment for dynamic systems formulated from the scalar quantities of kinetic energy T, potential energy U, and work W. The method is not confined to any specific coordinate system and results in a set of simultaneous differential equations of motion in terms of independent generalized coordinates.

Lagrange's equation is presented here as

$$(d/dt)(\partial T/\partial \dot{q}_i) - \partial T/\partial q_i + \partial U/\partial q_i = Q_i \qquad (59)$$

where Q_i is the generalized force. The generalized coordinates q_i are independent coordinates equal in number to the DOF of the system.

EXAMPLE. In Fig. 16. $q_1 = x$ and $q_2 = \theta$ are generalized coordinates completely defining the motion of the system. The velocities of the two masses are

$$v_1^2 = \dot{q}_1^2$$
$$v_2^2 = (\dot{q}_1 + l\dot{q}_2 \cos q_2)^2 + (l\dot{q}_2 \sin q_2)^2$$

and the kinetic energy becomes

$$T = \tfrac{1}{2}m_1\dot{q}_1^2 + \tfrac{1}{2}m_2\left[(\dot{q}_1 + l\dot{q}_2 \cos q_2)^2 + (l\dot{q}_2 \sin q_2)^2\right]$$

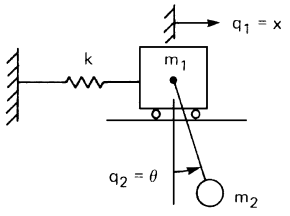


FIGURE 16 System with generalized coordinates q_1 and q_2.

Letting the potential energy be equal to zero along the horizontal line through m_1, the equation for U becomes

$$U = \tfrac{1}{2}kq_1^2 - m_2 gl \cos q_2$$

It is seen from these equations that T is a function of both $\dot{q}_i$ and q_i, whereas U is a function only of the q_i. Their substitution into Lagrange's equation results in a set of nonlinear differential equations in q_1 and q_2:

$$(m_1 + m_2)\ddot{q}_1 + m_2 l\left(\ddot{q}_2 \cos q_2 - q_2^2 \sin q_2\right) + kq_1 = 0$$
$$m_2 l^2 \ddot{q}_2 + m_2 l\ddot{q}_1 \cos q_2 + m_2 gl \sin q_2 = 0$$

If small angles are assumed for q_2 the linearized equations of motion become

$$(m_1 + m_2)\ddot{q}_1 + m_2 l\ddot{q}_2 + kq_1 = 0$$
$$m_2 l(\ddot{q}_1 + l\ddot{q}_2) + m_2 gl q_2 = 0$$

Generalized coordinates are independent coordinates equal in number to the DOF of the system. If, in the previous example, rectangular coordinates x_1, y_1 were chosen for the pendulum mass instead of the generalized coordinate $q_2 = \theta$, there would be one coordinate more than the required two coordinates for the problem. However, the rectangular coordinates x_1, y_1 are not independent and are related by the constraint equation $x_1^2 + y_1^2 = l^2$. Thus, one of the rectangular coordinates can be eliminated from the constraint equation, leaving only two independent coordinates as generalized coordinates.

In structural problems related coordinates are often encountered. However, the excess coordinates can be eliminated from the constraint equations, leaving the required number of coordinates as generalized coordinates with which to formulate the kinetic and potential energy expressions for the Lagrange's equations.

B. Mode Summation

Engineering structures and machines are generally composed of beams, columns, plates, shells, and other continuously distributed elements, each with an infinite number of DOF. The mode summation method enables one to analyze such systems as systems with a finite number of DOF.

In this method, the displacement of each component of the structure is represented by the sum of mode shapes $\phi_i(x)$. If these modes are normal modes, considerable simplification results due to orthogonality. Writing the deflection as

$$y(x, t) = \sum_i \phi_i(x) q_i(t) \qquad (60)$$

one can determine the kinetic energy, the potential energy, and the work done by external forces. These quantities are

then substituted into Lagrange's equation to establish the equations of motion for the system.

In terms of the generalized coordinates q_i, the following quantities are needed:

1. *Kinetic energy*

$$T = \frac{1}{2}\int \dot{y}^2(x,t)m(x)\,dx = \frac{1}{2}\sum_i\sum_j \dot{q}_i\dot{q}_j \int \phi_i\phi_j m(x)\,dx = \frac{1}{2}\sum_i\sum_j M_{ij}\dot{q}_i\dot{q}_j = \frac{1}{2}\dot{q}_i^{\mathrm{T}}[M_{ij}]\dot{q}_j \tag{61}$$

where $M_{ij} = \int \phi_i\phi_j m(x)\,dx$ is the generalized mass.

2. *Potential energy*

$$U = \frac{1}{2}\sum_i\sum_j K_{ij}q_iq_j = \frac{1}{2}q_i^{\mathrm{T}}[K_{ij}]q_j \tag{62}$$

The generalized stiffness K_{ij} depends on the type of elastic structure. For a uniform beam

$$K_{ij} = \int EI\phi_i''\phi_j''\,dx$$

3. *Work term*. The work term is established from the work done by the external forces due to a virtual displacement δq_i of the generalized coordinates:

$$\delta W = \int f(x,t)\left(\sum_i \phi_i\,\delta q_i\right)dx = \sum_i \delta q_i \int f(x,t)\phi_i(x)\,dx = \sum_i Q_i\,\delta q_i$$

The term

$$Q_i = \int f(x,t)\phi_i(x)\,dx \tag{63}$$

is called the generalized force.

These quantities can now be substituted into Lagrange's equation to establish the finite set of equations for the motion of the system. When $\phi_i(x)$ are normal modes; M_{ij} and K_{ij} become diagonal matrices, which leads to a set of uncoupled equations.

VI. APPROXIMATE AND NUMERICAL METHODS

Calculations for the natural frequencies and mode shapes of systems with many DOF are generally long and laborious, requiring the use of computers. In many cases, only a few of the lower modes are of interest, and simple and approximate methods can be used.

A. Rayleigh Quotient

When only the lowest fundamental frequency of a system is desired, the Rayleigh energy method offers a simple approach. The method can be applied to the multimass or distributed system.

We have found that, for a conservative system, the maximum value of the kinetic and potential energies can be equated. For a multimass or a distributed mass system this requires an assumption as to the displacement shape of the fundamental mode.

EXAMPLE. A continuously distributed system is often modeled by several lumped masses $m_1, m_2, m_3, \ldots$. To approximate the dynamic deflection, assume that the displacements of the masses are $y_1, y_2, y_3, \ldots$, produced by the static loads $m_1g, m_2g, m_3g, \ldots$. The maximum potential energy, stored as strain energy, is equal to the work done by the static loads of the masses:

$$U_{\max} = \tfrac{1}{2}g(m_1y_1 + m_2y_2 + m_3y_3 + \cdots)$$

With harmonic motion of frequency ω, the velocity of each mass is ωy_i, and the maximum kinetic energy is

$$T_{\max} = \tfrac{1}{2}\omega^2\left(m_1y_1^2 + m_2y_2^2 + m_3y_3^2 + \cdots\right)$$

Equating the two energies, we obtain the equation for the natural frequency known as the Rayleigh quotient:

$$\omega^2 = \frac{g(m_1y_1 + m_2y_2 + m_3y_3 + \cdots)}{\left(m_1y_1^2 + m_2y_2^2 + m_3y_3^2 + \cdots\right)} \tag{64}$$

Here we have assumed a deflection based on the static load of the masses. Since the deviation of the assumed curve from the exact dynamic curve can be thought of as constraints or added stiffness, the natural frequency calculated is slightly higher than the exact value. However, the method is somewhat insensitive to small deviations of the assumed curve and results in a fairly accurate value of the fundamental frequency.

Rayleigh's quotient can also be expressed in matrix form. Letting $\{x\}$ represent the assumed deflection vector, the potential and kinetic energies are

$$U_{\max} = \tfrac{1}{2}\{x\}^{\mathrm{T}}[K]\{x\}$$

$$T_{\max} = \tfrac{1}{2}\omega^2\{x\}^{\mathrm{T}}[M]\{x\}$$

Equating the two, the equation for the fundamental frequency in matrix form becomes

$$\omega^2 = \{x\}^{\mathrm{T}}[K]\{x\}/\{x\}^{\mathrm{T}}[M]\{x\} \tag{65}$$

B. Rayleigh–Ritz Method

Ritz extended the Rayleigh method to give a more accurate value of the fundamental frequency in addition to providing an approximation to the higher natural frequencies and their mode shapes. The Rayleigh–Ritz method starts by computing the fundamental frequency using

Rayleigh's equation

$$\omega^2 = U_{\max} \big/ T^*_{\max} \tag{66}$$

where the kinetic energy is given by

$$T = \omega^2 T^*$$

The assumed deflection in the Rayleigh equation is considered to be the sum of several mode functions multiplied by constants

$$y(x) = c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_m\phi_m(x) \tag{67}$$

The functions $\phi_i(x)$ are any admissible functions satisfying the geometric boundary conditions of the problem.

With $U_{\max}$ and $T^*_{\max}$ expressed in terms of $\phi_i(x)$ and c_i, the ω^2 is then minimized by differentiating with respect to each of the c_i. The result is n algebraic equation in ω^2, which in matrix notation is of the form

$$[f(\omega^2)]\{c_i\} = 0$$

The determinant of this equation yields the n natural frequencies of the system, and the corresponding mode shapes determined by the values of the c_i are found by substituting each ω^2 into the above equation.

The success of this method depends on the choice of the shape functions $\phi_i(x)$, which calls for some experience on the part of the analyst.

C. Method of Matrix Iteration

The equations of motion, previously formulated in terms of stiffness, can also be expressed in terms of flexibility. The flexibility influence coefficient a_{ij} is defined as the deflection at i due to a unit load at j. Thus, in Fig. 17, a_{12} is the deflection at 1 due to a unit force at 2. Similarly a_{32} is the deflection at 3 due to the same loading.

With several forces acting, the principle of superposition enables one to write

$$\begin{aligned} y_1 &= a_{11}f_1 + a_{12}f_2 + a_{13}f_3 \\ y_2 &= a_{21}f_1 + a_{22}f_2 + a_{23}f_3 \\ y_3 &= a_{31}f_1 + a_{32}f_3 + a_{33}f_3 \end{aligned}$$

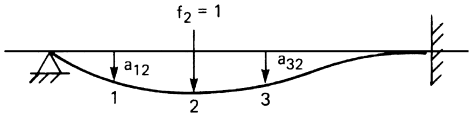


FIGURE 17 Flexibility influence coefficients.

which in matrix notation becomes

$$\{y\} = [a_{ij}]\{f\} \tag{68}$$

The square matrix $[a_{ij}]$ is here the flexibility matrix. Since $\{f\} = [k]\{y\}$, the flexibility matrix must be the inverse of the stiffness matrix.

For a system vibrating in one of the normal modes, the loading is equal to the inertia loads,

$$f_i = -m_i\ddot{y}_i = m_i\omega^2 y_i$$

Thus, Eq. (68) can be written

$$\{y\} = [a]\{m\omega^2 y\}$$

which written out becomes

$$\begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \omega^2 \begin{bmatrix} a_{11}m_1 & a_{12}m_2 & a_{13}m_3 \\ a_{21}m_1 & a_{22}m_2 & a_{23}m_3 \\ a_{31}m_1 & a_{32}m_2 & a_{33}m_3 \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} \tag{69}$$

This equation is suitable for the matrix iteration procedure, which converges to the lowest mode corresponding to the fundamental frequency.

Starting out with an assumed deflection for the right column and performing the indicated multiplication, a new deflection column is obtained. Normalizing by letting one of the deflections be unity, the procedure is repeated any number of times until the normalized deflection converges to a steady configuration, which is the fundamental mode of vibration. The normalization will also result in a value of ω^2 for the fundamental frequency.

Higher modes can be found, provided that the lower modes are eliminated from the assumed deflection. This is accomplished by assuming the trial deflection to be the sum of normal modes multiplied by constants and systematically eliminating the lower modes by the use of orthogonality. The method, which will not be discussed here, is called sweeping because it results in a sweeping matrix to sweep out the lower modes.

D. Holzer Method

Holzer proposed a simple numerical procedure for the calculation of the natural frequencies and mode shapes of any multi-DOF torsional system. The method can also be used for the translational vibration of a lumped-mass-spring system.

For the torsional system, the calculations are started by assuming a unit torsional amplitude at one end; after a trial frequency ω is chosen, the torques and amplitudes are progressively calculated to the other end of the system. Depending on the boundary conditions, the torque or the angular displacement at the boundary is plotted against the frequency ω. If the boundary is free, the frequencies that result in zero torque are the natural frequencies of the

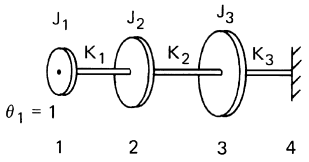


FIGURE 18 Torsional system of three DOF.

system. If the boundary is fixed, the frequencies that result in zero displacement are the natural frequencies.

EXAMPLE. The three-DOF torsional system of Fig. 18 is free at 1 and fixed at 4. A frequency ω having been chosen, the torque required on disk 1 for $\theta_1 = 1$ is $\omega^2 J_1$. This torque acting on shaft 1 will twist it by

$$\omega^2 J_1/K_1 = 1 - \theta_2$$

or

$$\theta_2 = 1 - \omega^2 J_1/K_1$$

The torque required on disk 2 to maintain the amplitude is $\omega^2 J_2\theta_2$, and shaft 2, supporting the sum of the torques of disk 1 and 2, will twist by

$$\theta_2 - \theta_3 = \left(\omega^2 J_1 + \omega^2 J_2\theta_2\right)/K_2$$

Again the torque of disk 3 to maintain the amplitude θ_3 is $\omega^2 J_3\theta_3$, and shaft 3 will twist by

$$\theta_3 - \theta_4 = \left(\omega^2 J_1 + \omega^2 J_2\theta_2 + \omega^2 J_3\theta_3\right)/K_3$$

Repeating the same calculation with another value of ω, θ_4 is calculated and plotted against the new value of ω. A plot of θ_4 against ω may then appear as in Fig. 19 with θ_4 passing through zero at three values of ω. These are the three natural frequencies of the system.

The mode shapes can be determined by calculating the values of θ_i from the above equations, using the ω's for the natural frequencies.

1. Transfer Matrix

In the Holzer method, the state of the deflection and torque at one station is transferred to the neighboring station, and

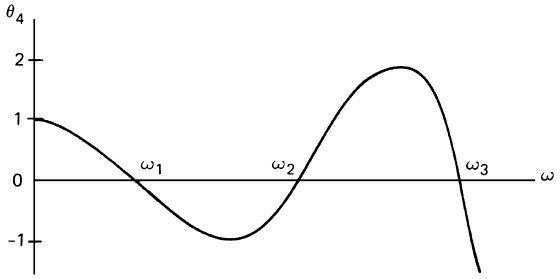


FIGURE 19 Natural frequencies of torsional system of Fig. 18.

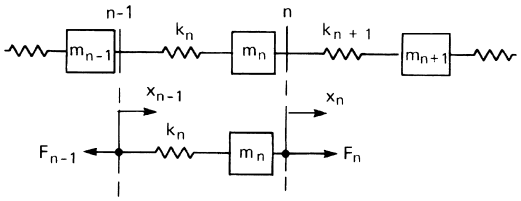


FIGURE 20 Element of transfer matrix.

the procedure is numerically carried out from one end of the system to the other. The transfer matrix method is a matrix systemization of the Holzer method. The method can also be applied to the linear spring-mass system and to beams and branched systems.

The existing state at any station is first defined by the state vector, which is a column matrix of the deflection and force. For the spring–mass system of Fig. 20 the stations are numbered with the spring and the mass to the right as the structural element. The state vector for this system is the deflection and force at $n-1\{\begin{smallmatrix} x \\ F \end{smallmatrix}\}_{n-1}$, which is to be related to $\{\begin{smallmatrix} x \\ F \end{smallmatrix}\}_n$.

Considering the spring k_n, the displacement at the two ends are x_n and x_{n-1}, and the force through it is F_{n-1}. The equation relating the two is

$$x_n = x_{n-1} + F_{n-1}/k_n$$

which can be written by the matrix equation

$$\begin{Bmatrix} x \\ \end{Bmatrix}_n = \begin{bmatrix} 1 & 1/k_n \\ & \end{bmatrix} \begin{Bmatrix} x \\ F \end{Bmatrix}_{n-1}$$

The forces on the two sides of m_n are F_n and F_{n-1} and the equation is

$$F_n = F_{n-1} - \omega^2 m_n x_n$$

Substituting for x_n from the first equation, we have

$$F_n = F_{n-1} - \omega^2 m_n (x_{n-1} + F_{n-1}/k_n)$$

or

$$\begin{Bmatrix} \\ F \end{Bmatrix}_n = \begin{bmatrix} & \\ -\omega^2 m_n & (1 - \omega^2 m_n/k_n) \end{bmatrix} \begin{Bmatrix} x \\ F \end{Bmatrix}_{n-1}$$

Putting the two matrix equations together, the desired result is

$$\begin{Bmatrix} x \\ F \end{Bmatrix}_n = \begin{bmatrix} 1 & 1/k \\ -\omega^2 m & (1 - \omega^2 m/k) \end{bmatrix}_n \begin{Bmatrix} x \\ F \end{Bmatrix}_{n-1}$$

which transfers the state vector at $n-1$ to the state vector at n. The square matrix above is called the transfer matrix for the nth element.

Starting with a numerical value for ω^2, the calculation can be progressively carried out from one end of the system to the other. Depending on the boundary conditions, either

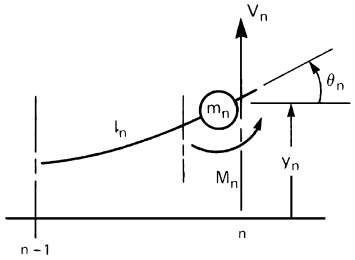


FIGURE 21 Beam element.

x_n or F_n at the far end can be plotted against ω^2, and the natural frequencies of the system are found when the boundary conditions are satisfied.

The procedure to be followed for the torsional system is identical to that of the linear spring–mass system, the state vector being the angular displacement θ and the torque T.

For the beam, the mass is again lumped at the right end, as shown in Fig. 21. The state vector will here contain four quantities $\{-V \;\; M \;\; \theta \;\; y\}^{\mathrm{T}}$, where V is the shear, M the bending moment, θ the slope of the beam, and y its lateral deflection.

The transfer matrix can be developed in two parts, one called the field matrix, related to the elastic element, and the other called the point matrix, related to the quantities on the two sides of the mass. The two are then put together for the transfer matrix, which is a 4×4 matrix. Again the numerical procedure starts with a chosen value of ω^2, and the boundary conditions must be satisfied for the determination of the natural frequencies.

E. Finite Difference Numerical Computation

When the differential equation cannot be integrated in closed form, numerical methods must be employed. This may occur when the system is nonlinear or if the system is excited by a force that cannot be expressed by simple analytic functions.

In the finite difference method for initial value problems, the continuous variable t is replaced by the discrete variable t_i. The differential equation is solved progressively in time increments $h = \Delta t$ starting from known initial conditions. With a sufficiently small time increment, an approximate solution of acceptable accuracy is obtainable.

In this section we discuss two finite difference methods. A discussion of the merits of the different finite difference methods such as accuracy, stability, and length of computation are beyond the scope of this article.

In the first method, the second-order differential equation for the viscously damped single DOFS

$$M\ddot{x} + C\dot{x} + Kx = F(t) \tag{70}$$

is solved directly by discretizing the derivatives using the central difference method. This method is developed from the Taylor expansion of x_{i+1} and x_{i-1} about the point i.

$$x_{i+1} = x_i + h\dot{x}_i + \frac{h^2}{2}\ddot{x}_i + \frac{h^3}{6}\dddot{x}_i + \cdots$$
$$x_{i-1} = x_i - h\dot{x}_i + \frac{h^2}{2}\ddot{x}_i - \frac{h^3}{6}\dddot{x}_i + \cdots \tag{71}$$

where the time interval is $h = \Delta t$. Subtracting and ignoring terms of order h^2 and higher, we obtain

$$\dot{x}_i \frac{1}{2h}(x_{i+1} - x_{i-1}). \tag{72}$$

Adding, we find

$$\ddot{x}_i = \frac{1}{h^2}(x_{i-1} - 2x_i + x_{i+1}). \tag{73}$$

Replacing the derivatives in Eq. (70) by the central differences the finite difference equation is given by

$$\frac{M}{h^2}[x_{i-1} - 2x_i + x_{i+1}] + \frac{C}{2h}[x_{i+1} - x_{i-1}] + Kx_i = F_i \tag{74}$$

where $x_i = x(t_i)$ and $F_i = F(t_i)$. Rearranging this equation yields the recurrence formula

$$x_{i+1} = \left[\frac{1}{\frac{M}{h^2} + \frac{C}{2h}}\right]\left[F_i + \left[\frac{2M}{h^2} - K\right]x_i + \left[\frac{C}{2h} - \frac{M}{h^2}\right]x_{i-1}\right]. \tag{75}$$

It allows us to compute the displacement of the mass at time t_{i+1}, x_{i+1}, if we know the displacements at time t_i and t_{i-1} and the external force F_i. This formula is not self-starting. In order to find x_1, we need x_0 which is given as an initial condition and x_{-1} which is not known but can be computed. The initial values x_0 and $\dot{x}_0$ are used to compute $\ddot{x}(0)$ from the differential equation

$$\ddot{x}_0 = \frac{1}{M}[F(0) - C\dot{x}_0 - Kx_0] \tag{76}$$

The value of x_{-1} is obtained by evaluating the backward Taylor expansion about $i = 0$ to get

$$x_{-1} = x_0 - h\dot{x}_0 + \frac{h^2}{2}\ddot{x}_0 \tag{77}$$

Equations (75–77) constitute the central difference method for the viscously damped single degree of freedom vibrating system.

The second finite difference method, known as the Runge-Kutta method, is also based on Taylor series expansions. The fourth-order Runge-Kutta method presented below matches the Taylor series expansion up to terms

of order h^4 without explicitly computing derivatives beyond the first. It does this by judiciously combining four different evaluations of the first derivative. This method is popular because it is self-starting, i.e. it only uses the initial conditions to compute x_1, and results in good accuracy.

Since the Runge-Kutta method approximates first derivatives, the second-order differential equation needs to be converted into a system of two first order equations. This means that the differential equation for the single degree of freedom viscously damped system

$$\ddot{x} = \frac{1}{M}[F(t) - Kx - C\dot{x}] \tag{78}$$

becomes the system of equations

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= \frac{1}{M}[F(t) - Kx - Cy] \end{aligned} \tag{79}$$

By defining

$$\vec{X} = \begin{pmatrix} x \\ y \end{pmatrix} \tag{80}$$

and

$$\vec{G} = \begin{pmatrix} y \\ \frac{1}{M}[F(t) - Kx - Cy] \end{pmatrix} \tag{81}$$

The fourth order Runge-Kutta method results in the following recurrence formula

$$\vec{X}_{i+1} = \vec{X}_i + \tfrac{1}{6}[\vec{K}_1 + 2\vec{K}_2 + 2\vec{K}_3 + \vec{K}_4] \tag{82}$$

where

$$\begin{aligned} \vec{K}_1 &= h\vec{G}(\vec{X}_i, t_i) \\ \vec{K}_2 &= h\vec{G}(\vec{X}_i + 0.5 * \vec{K}_1, t_i + 0.5 * h) \\ \vec{K}_3 &= h\vec{G}(\vec{X}_i + 0.5 * \vec{K}_2, t_i + 0.5 * h) \\ \vec{K}_4 &= h\vec{G}(\vec{X}_i + \vec{K}_3, t_{i+1}) \end{aligned} \tag{83}$$

Equations (82) and (83) constitute the fourth-order Runge-Kutta method.

VII. CONCLUSIONS

The subject of vibration covers a wide area with many interesting analytic techniques and methods of computation. Obviously many of these areas cannot be presented comprehensively in a short summary article such as this and have thus been omitted.

The digital computer has made possible the solution of problems that previously defied computation and has revolutionalized our treatment of these problems. It has introduced new concepts of analysis such as the finite element approach, which is capable of solving very large structural problems.

Two general areas of vibration that differ markedly from the subjects presented here should be mentioned briefly. The first area is the vibration of nonlinear systems. Its most important difference arises from the fact that the principle of superposition, which plays a major role in the vibration theories of linear systems, no longer applies to the nonlinear system. Mathematical difficulties are encountered in solving nonlinear differential equations. However, there is no particular difficulty in obtaining numerical solutions with the digital computer.

The second area that requires a different approach is that of random vibrations. These are vibrations produced by forces varying in a random manner, which can be defined only by probability and statistical terms. For example, air gusts encountered by an airplane in flight can be defined only in terms of statistical averages and probability of encounter. Obviously, the response of a structure to such random excitation is also random and must be defined in terms of statistics and probability. Any one of these areas would require extensive study.

SEE ALSO THE FOLLOWING ARTICLES

ELASTICITY • MECHANICS, CLASSICAL • NONLINEAR DYNAMICS • NUMERICAL ANALYSIS • WAVE PHENOMENA

BIBLIOGRAPHY

Bathe, E. C., and Wilson, E. L. (1976). "Numerical Methods in Finite Element Analysis," Prentice-Hall, Englewood Cliffs. New Jersey.

Benaroya, H. (1998). "Mechanical Vibration: Analyasis, Uncertainties, and Control," Prentice-Hall, Eaglewood Cliffs, New Jersey.

Craig, R., Jr. (1981). "Structural Dynamics," Wiley, New York.

Gerald, C., and Wheatley, P. (1997) "Applied Numerical Analysis," Addison-Wesley, Reading, Massachusetts.

Meirovitch, L. (1967). "Analytical Methods in Vibrations," Macmillan, New York.

Meirovitch, L. (1980). "Computational Methods in Structural Dynamics," Sijthoff & Noordhoff, Rockville, Maryland.

Rayleigh, L. (1945). "The Theory of Sound," Vol. 1, Dover, New York.

Thomson, W. T., and Dahleh M. (1998). "Theory of Vibration with Applications," Prentice-Hall, New Jersey.



Wave Phenomena

Norman Bleistein
Colorado School of Mines

I. Waves in One Dimension: Fundamental Concepts
II. Waves in Higher Dimensions

GLOSSARY

Amplitude Local peak amplitude of a wave form; function A.

Dot product $\boldsymbol{k} \cdot \boldsymbol{x} = k_1 x_1 + k_2 x_2 + k_3 x_3$ in three dimensions, $k_1 x_1 + k_2 x_2$ in two dimensions.

Frequency Temporal (local) rate at which a wave repeats its fundamental form; ω in the units of radians/second, $f = \omega/2\pi$ in the units cycles/second or hertz.

Group speed Speed at which energy propagates; in one dimension, $|d\omega/dk|$; in higher dimensions, $|\nabla_k \omega(\boldsymbol{k})|$.

Group velocity Velocity vector that describes the magnitude and the direction of the propagation of energy; in one dimension, $d\omega/dk$; in higher dimension, $\nabla_k \omega(\boldsymbol{k})$.

Incident wave Wave on one side of a surface whose propagation is toward the surface.

Period Elapsed time for one cycle of a wave; $2\pi/\omega = 1/f$.

Phase The function $kx - \omega t$ or $\boldsymbol{k} \cdot \boldsymbol{x} - \omega t$ in the waveforms above.

Phase speed Speed at which crests of a wave propagate; in one dimension, $|\omega/k|$; in higher dimensions, $|\omega|/k$, with k being the magnitude of the wave vector, $\boldsymbol{k}$.

Phase velocity Both speed and direction at which wave crests propagate; $\hat{k}\omega(\boldsymbol{k})/k$.

Rays Trajectories along which the constituent components of a wave—wave vector, frequency, phase, and energy—propagate.

Reflected wave Wave arising at a surface of discontinuity (interface↔reflector) in the propagation parameters of a medium; this wave propagates on the same side of the interface as the incident wave.

Refracted wave Wave arising at a surface of discontinuity (interface) in the propagation parameters of a medium; this wave propagates on the opposite side of the interface from the incident wave, and the propagation direction of this wave and the propagation direction of the incident wave satisfy Snell's law.

Snell's law Law relating the directions of incidence and refraction of a wave at an interface. [See Eq. (79).]

Stationary phase Method for obtaining an approximation (asymptotic expansion) of an integral with an oscillatory factor, such as a Fourier superposition integral.

Wavelength Fundamental length scale over which a wave repeats itself; $2\pi/k$.

Wavenumber Spatial rate at which cycles of a wave occur; coefficient k; for the higher dimensional case, k is the magnitude of $\boldsymbol{k}$.

Wave vector $\boldsymbol{k} = (k_1, k_2, k_3)$ in three dimensions; (k_1, k_2) in two dimensions; the notation (k_x, k_y, k_z) is also used.

THE PHENOMENON of wave motion is the primary mechanism by which a disturbance transfers energy over a distance in a medium. The propagation of this energy is thought of as being wavelike when it can be characterized

by some feature (e.g., a crest) that is at least partially preserved as recognizable during the propagation over a distance or time interval. The most common wave phenomena are acoustic (sound), elastic (seismic), electromagnetic (light, radio, or television), or gravitational (surface water) waves; there are many others. Certain features of the propagation of waves are common to all wave phenomena, no matter what the medium.

This article is a partial description of the broad class of common features of wave phenomena as seen in their mathematical description. Where it is necessary to distinguish between linear and nonlinear waves, this discussion is further limited to the former. Even a textbook-sized discussion would inevitably omit some common features of waves—linear and nonlinear—because of the breadth of the subject. This, then, is one author's choice of a fundamental subset of common features of linear wave phenomena.

The discussion starts with one-dimensional wave propagation. We start with definitions of the features of a single sinusoidal wave—amplitude, wavelength, wavenumber, period, frequency. We then proceed to a simple superposition of two waves to introduce the distinction between phase speed/velocity and group speed/velocity.

These simple ideas then become a point of departure for the discussion of Fourier superposition. This is a powerful tool for deriving analytical representations of solutions of wave equations in homogeneous media. It further has application to provide exact representation in some cases of heterogeneous media and, beyond that, it provides approximate representations of wave fields in an even larger class of heterogeneous media.

However, when synthesizing waves over a continuum of wavenumbers, the identification of phase velocity and group velocity is obscured by the representation. In order to recapture those features of wave propagation, the method of stationary phase is introduced. It is shown that this approximation of the wave provides a conceptually simplified interpretation of the more complicated Fourier synthesis. In this simplified representation, the phase and group velocities of the individual elements of the Fourier synthesis again become apparent, but this representation is an approximation of the original integral. We present a numerical example to demonstrate the reliability of this approximation under appropriate dimensionless constraints on the physical parameters of the wave being represented.

The same development is repeated for higher dimensional wave propagation. In this case, there are additional features due to the dimensionality: The wavenumber is replaced by a wave vector; directionality plays an important role in the identification of phase and group velocities. Interestingly, these two velocities need not coalign.

Again, Fourier synthesis provides a means for describing more complicated waves and a *multidimensional* stationary phase provides a means of approximating those waves that admits simpler interpretation in terms of wave packets propagating with their own group velocity, while elements at specific wave vectors within the group travel with their own individual phase velocity.

The article closes with discussion of reflection and refraction of a three-dimensional plane wave by a planar reflector.

I. WAVES IN ONE DIMENSION: FUNDAMENTAL CONCEPTS

As a specific example to picture in our minds, let us suppose that we are describing the vertical displacement of points on a straight line (a string) as a function of transverse location (x) on the line and time (t). We shall denote the vertical displacement by $u(x, t)$. As a simple example of that displacement, let us suppose that u is given by

$$u(x, t) = A\cos(kx - \omega t). \tag{1}$$

In this equation, A, k, and ω are constants; for now, they are all positive constants. (Note that we could have as easily begun our discussion using a sine function instead of a cosine function.)

A. Amplitude, Phase, Wavelength, Wavenumber, Period, and Frequency

For each fixed value of t, the graph of $u(x, t)$ in the (x, u)-plane is a cosine function of maximum height A called the amplitude of the wave. The argument of the cosine function $[kx - \omega t]$ is called the *phase* of the wave. The peaks or crests of the cosine function, that is, the points where $u(x, t) = A$, occur whenever

$$kx = 2n\pi + \omega t, \quad n = \ldots, -2, -1, 0, 1, 2, \ldots \tag{2}$$

The peaks are separated by a distance over which kx increases by 2π, namely a distance

$$\lambda = 2\pi/k \tag{3}$$

called the *wavelength* of the wave represented by $u(x, t)$ (Fig. 1). The constant k is called the *wavenumber*.

For fixed x and variable t, the graph of $u(x, t)$ in a (t, u)-plane is analogous to what we have just described. The amplitude of u is again given by A, but now the peaks of the cosine function at fixed x occur at the times

$$\omega t = 2m\pi + kx, \quad m = \ldots, -2, -1, 0, 1, 2, \ldots \tag{4}$$

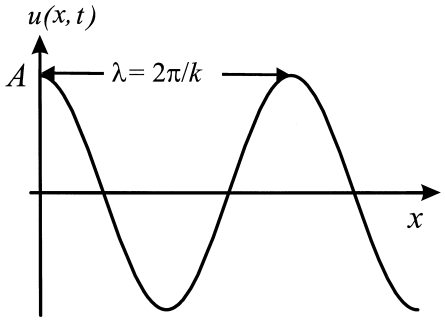


FIGURE 1 The wave of Eq. (1) for fixed t.

The elapsed time between peaks at fixed x is such that the increment in ωt is equal to 2π, given by a time

$$T = 2\pi/\omega \tag{5}$$

called the *period* of the wave motion. The constant ω is called the *frequency* (Fig. 2).

Of course, we can look at the wave as a function of x and t simultaneously; see Fig. 3. For this example, we have chosen $\omega = 2k$. This manifests itself as an apparent compression of the wave crests in the t-direction as compared to the density of the wave crests in the x-direction.

Now think of a vertical plane parallel to the x–u-plane—a constant t-plane. This provides a snapshot, such as the one in Fig. 1. Now consider moving that plane in the positive t-direction. From the figure, it should be apparent that each wave crest, each wave trough—in fact, every point of constant phase on the wave—moves in the positive x-direction, increasing x. This is a manifestation of positive phase speed, a subject of the next section.

The units of the phase function in Eq. (1) are *radians*. Therefore, the units of the wavenumber k are radians per unit length, while the units of the frequency ω are radians per unit time. Because there are 2π radians per period or cycle, it is sometimes more convenient to use units of frequency and wavenumber that are scaled by 2π, which is the number of radians in one period or cycle. Thus, the new variables have the dimensions of cycles per unit time

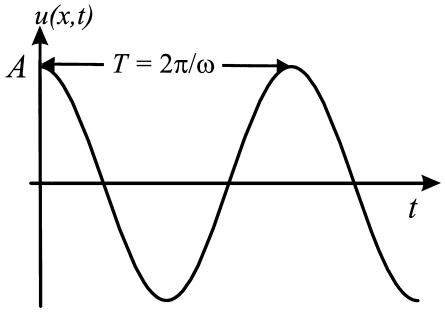


FIGURE 2 The wave of Eq. (1) for fixed x.

or cycles per unit length. These variables are often denoted by f and f_x, defined by

$$\omega = 2\pi f \qquad \text{and} \qquad k = 2\pi f_x, \tag{6}$$

respectively. The units of f are reciprocal time, often referred to as cycles per unit time, and the units of f_x are reciprocal length, referred to as cycles per unit length. When the time unit is seconds, the units of f are called hertz (Hz). In these units, the temporal period and the frequency are reciprocals of one another, as are the spatial period and wave number, now often referred to as the spatial frequency.

B. Phase Speed and Group Speed

Having examined the function $u(x,\ t)$ for both fixed t and fixed x, we are now prepared to consider u when both x and t are allowed to vary. In particular, let us consider the graph in the (x, u)-plane. The peaks of u, as well as all the points of constant phase, and hence constant u, will move or propagate as time progresses. The rate (v_ϕ) at which a point of constant phase will move is readily determined by setting the phase equal to a constant and differentiating that relationship with respect to t:

$$kx - \omega t = \text{const.}, \qquad v_\phi = dx/dt = \omega/k \tag{7}$$

Thus, we see that the points of constant phase move with the speed ω/k, called the *phase speed*. When ω and k have the same sign, this motion is to the right; when ω and k have opposite signs, the motion is to the left.

The wave $u(x, t)$ defined by Eq. (1) is *periodic*, having exactly the same shape in every interval whose length is given by the wavelength λ. It is also periodic in t, having the same shape in every temporal interval given by the period T. In reality, no wave can be periodic over all space and time. However, many wave phenomena are periodic on intervals of sufficient length (many multiples of λ) and/or for intervals of sufficient time (many multiples of T) to be considered periodic for all practical purposes. (A simple example would be alternating current in a transmission line or waveguide.) Indeed, the transmission of information in an otherwise periodic wave depends on local variations in amplitude (amplitude modulation) or phase (frequency modulation).

In many cases, the phase velocity v_ϕ varies with ω and k. Typically, the physics of a particular problem and its attendant mathematical model impose a relationship between k and ω, called a *dispersion relation*

$$\omega = \omega(k) \tag{8}$$

Except in the special case in which $\omega = ck$, with c independent of k, different frequencies will propagate at different

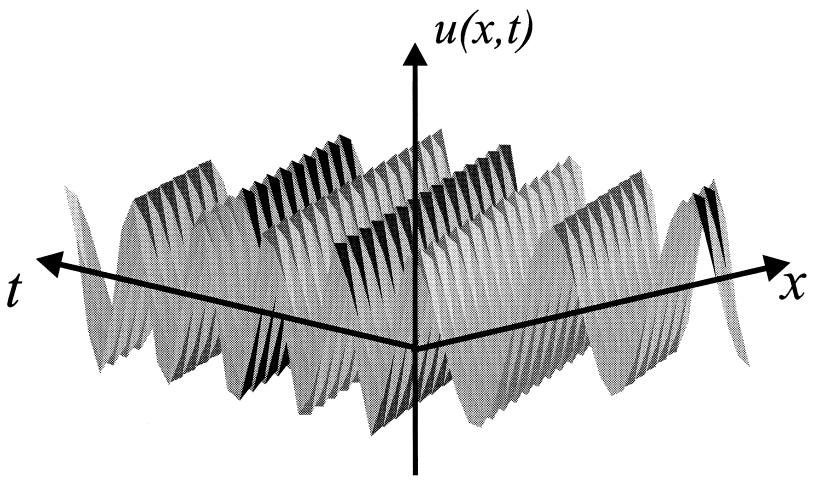


FIGURE 3 A space-time image of the wave of Eq. (1).

speeds determined by the dispersion relation and the definition of v_ϕ in Eq. (7).

We next consider waves of two nearby frequencies and the same amplitude and ask how the composite wave, which is the sum of the two, will propagate. Thus, let us introduce the function

$$u(x,t) = A[\cos(k^+x - \omega^+t) + \cos(k^-x - \omega^-t)] \quad (9)$$

In this equation, we have used $k^\pm$ and $\omega^\pm$ as shorthand notations for

$$\begin{aligned} k^\pm &= \bar{k} \pm \Delta k; \qquad \omega^\pm = \bar{\omega} \pm \Delta\omega \\ &\approx \bar{\omega} \pm (d\omega/dk)\Delta k \qquad (10) \\ \bar{k} &= (k^+ + k^-)/2;\ \bar{\omega} = \omega(\bar{k}) \end{aligned}$$

By using the appropriate trigonometric identity, we can rewrite the sum of cosine functions in Eq. (9) as a product of cosines:

$$u(x,\ t) = 2A\cos(\Delta kx - \Delta\omega t)\cos(\bar{k}x - \bar{\omega}t) \quad (11)$$

Implicit in our notation is the assumption that Δk is much smaller than $\bar{k}$, so that the wavelength $2\pi/\Delta k$ associated with the first cosine factor in this equation is much larger than the wavelength $2\pi/k$ associated with the second cosine factor. Thus, the first cosine factor acts as a slowly varying amplitude modulator, varying the amplitude $2A$, which is the sum of the two amplitudes of the constituent waves of $u(x,\ t)$. The wave of average wavenumber $\bar{k}$ and average frequency $\bar{\omega}$ travels through the envelope at its phase speed $v_\phi = \bar{\omega}/\bar{k}$, while the envelope itself moves at its own speed, associated with the differentials Δk and $\Delta\omega$,

$$v_g = \Delta\omega/\Delta k \approx d\omega/dk|_{k=\bar{k}} \quad (12)$$

known as the *group speed*.

Suppose that v_ϕ and v_g are both positive. When $v_\phi > v_g$, the crests moving at the phase speed move forward through each wavelength of the packet created by the modulator of the amplitude; when $v_\phi < v_g$, the crests move backward through the packet. The former case—or more precisely, $v_\phi \geq v_g$—is more typical, with v_g having an upper bound, the characteristic speed of the medium (e.g., sound speed, light speed) through which the wave propagates, and v_ϕ having the characteristic speed as a lower bound.

In Fig. 4, we show a sum of the two waves of Eq. (9). They are of unit amplitude with $\bar{k} = \pi$, $\Delta k = 0.05k$, and $t = 1$. Further, $\omega(k) = \sqrt{k^2 + \pi^2}$. The x-range here is 40 units in the given length scale. Thus, we see 20 cycles of the high-fequency wave over this range. On the other hand, the sum of the waves is equal to zero at $x = 20$, where $\Delta kx = 0.05\pi \times 20 = \pi$, and the arguments of the two cosine functions are out of phase by π, making the sum equal to zero. In Fig. 5, we show the same wave at $t = 4.353$. The peaks and the zero of the envelope have moved forward and the peaks of the fast cycles do not occupy the same positions in the envelope. Actually, they have moved forward, as well.

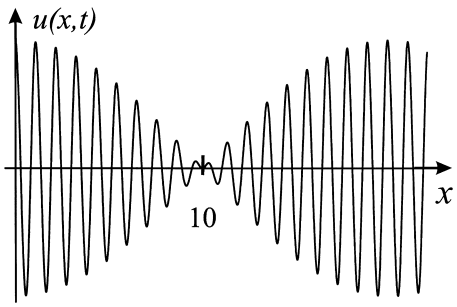


FIGURE 4 An example of Eq. (9) at $t = 0$.

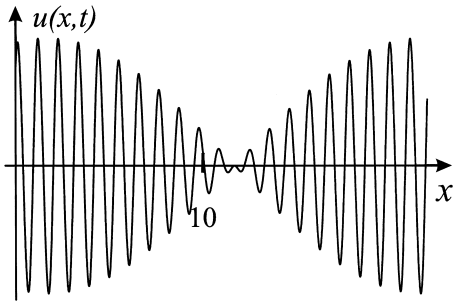


FIGURE 5 The same example of Eq. (9) as in Fig. 4 at $t = 4.353$.

An important feature of the group speed is that the *energy* of the wave residing in the wavenumbers near $\bar{k}$ will propagate at this speed. Thus, if a localized disturbance is created, it is the group speed that will determine how much time will elapse before this portion of the disturbance is observed at a distance.

C. Fourier Superposition

These ideas extend in a natural way to the Fourier superposition of waves expressed as

$$u(x,\ t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} A(k)\exp\{i[kx - \omega(k)t]\}dk \qquad (13)$$

In this equation, we think of $A(k)\,dk/2\pi$ as the amplitude of a wave with wavenumber k and frequency $\omega(k)$. The integration (summation) is then a superposition over all values of k. The values of k for which $A(k)$ are nonzero are called the *spectrum* of the wave $u(x,\ t)$. The product $A(k)\,dk$ must have the same dimensions as u itself. Thus, $A(k)$ must have the dimensions of u/unit-length of k; that is, $A(k)$ is a density, called the *spectral density* of the wave $u(x,\ t)$.

We have used the complex exponential for our Fourier superposition, but we assume that the amplitude function $A(k)$ is such that the resulting integral is real. For example, suppose that $\omega(k)$ were an odd function of k so that negative frequencies yielded an exponential function for negative k that is the complex conjugate of its values for positive k. Then, when the real part of A $\mathrm{Re}\{A\}$ is an even function of k and the imaginary part of A $\mathrm{Im}\{A\}$ is an odd function of k, $u(x,\ t)$ would be real. Under other assumptions on $\omega(k)$, other constraints on A would make the resulting integral real. Alternatively, we could simply require that u be defined by the real part of the integral on the right.

D. Stationary Phase Formula

Let us suppose in Eq. (13) that $A(k)$ is nonzero only for values of $|k|$ larger than some minimum value, say k_0. We define $\omega_0 = |\omega(k_0)|$ as the associated frequency. We assume that $|\omega(k)| \geq \omega_0$ whenever $|k| \geq k_0$. We then rewrite the exponent in Eq. (13) as

$$kx - \omega(k)t = \omega_0 t\,[kx/(\omega_0 t) - \omega(k)/\omega_0] \qquad (14)$$

In this form, we may think of $\omega_0 t$ as playing the role of a dimensionless parameter to be denoted by Λ (see below) and the expression $kx/(\omega_0 t) - \omega(k)/\omega_0$ as a dimensionless phase function with independent variable k. We could as well make the independent variable dimensionless by scaling k by k_0; that is, $k/k_0 = \eta$. Later, we will describe the analysis of integrals such as Eq. (13) in terms of such dimensionless variables.

In practice, the parameter Λ is often large. We offer the following interpretation of this requirement. Let us denoted by T_0 the period associated with the minimum frequency ω_0; that is, $T_0 = 2\pi/\omega_0$. Then $2\pi t/T_0$ must be large. That is, the observation time multiplied by 2π must be "many" periods at the minimum frequency. Most often, this requirement is stated in a form that puts the burden on the frequency rather than the time. That is, the frequency is such as to make $\omega_0 t$ large. Thus, we may think of large Λ as characterizing *high frequency*. Although we have described this as being "many" periods, note that the factor of 2π in the expression $2\pi t/T_0$ provides some help in this matter. In practice, one often finds that

$$2\pi t/T_0 \geq \pi \Rightarrow t \geq T_0/2$$

is good enough! That is, the asymptotic approximation that is described below provides a "reasonably" accurate description of the integral, Eq. (14), for times beyond a half period.

By scaling out the factor $k_0 x$, we could have obtained an interpretation in terms of propagation over many units of inverse wavenumber instead of many periods. In either case, we must only require that, after scaling, the dimensionless derivatives should be bounded and should not be comparable in magnitude to the dimensionless large parameter Λ. That is, the remaining phase function should be "slowly varying" when compared to Λ.

In this limit we can approximate the integral in Eq. (13) by the method of stationary phase. We state the basic result for one-dimensional integrals here. In the following section, the result for multidimensional integrals will be presented. Suppose that

$$I = \int f(\eta)\exp\{i\Lambda\Phi(\eta)\}d\eta \qquad (15)$$

with Λ being a large parameter, in practice at least 3 or π, as was used in the earlier discussion. Then the value of the integral will be dominated by its contributions from the neighborhood of certain points, say $\eta_j,\ j = 1, 2, \ldots, n$,

called stationary points, where the first derivative vanishes:

$$d\Phi/d\eta = 0, \quad \eta = \eta_j, \qquad j = 1, 2 \ldots, n \tag{16}$$

When the second derivative at the stationary point does not vanish, the point is called a *simple stationary point*. In practice, it is most often the case that the stationary points are simple. Of course, the case of higher order stationary points (where a higher order derivative is the first nonvanishing derivative at the stationary point) occurs as well and leads to a rich theory of wave phenomena beyond the scope of the present discussion. We proceed under the assumption that the stationary points are simple. In this case, the integral I defined by Eq. (15) is approximated by

$$I \sim \sum_{j=1}^{n} \sqrt{\frac{2\pi}{|\Lambda \Phi''(\eta_j)|}}\, f(\eta_j) \exp[i\Lambda\Phi(\eta_j) + i(\pi/4)\mathrm{sgn}(\Lambda)\mathrm{sgn}(\Phi''(\eta_j))] \tag{17}$$

This is the *stationary phase formula* for the case of a simple stationary point. In this equation, we have used prime ($'$) to denote differentiation with respect to η. The notation $\mathrm{sgn}(\Lambda)$ means "sign of Λ." The symbol "$\sim$" is to be read as "is asymptotically equal to." It means that the error approaches zero more rapidly than the terms of the sum, that is, more rapidly than a constant over $\sqrt{|\Lambda|}$, as $\Lambda \to \infty$. Usually the error is bounded by a constant over $|\Lambda|$ or a constant over $|\Lambda|^{3/2}$.

Despite the formal statement addressing the error only in the limit as $|\Lambda| \to \infty$, we repeat that in practice $|\Lambda|$ greater than 3 or π—use whichever is convenient—would seem to suffice. For example, when this asymptotic approximation is used to estimate the zeroth-order Hankel function of the first kind for its argument equal to 3 that is, $H_0^{(1)}(3)$, the error turns out to be only about 6%, sufficiently small for a qualitative understanding of how the function in question behaves and even adequate for purposes of modeling of real-world wave phenomena.

The method of stationary phase quantifies the following qualitative ideas about the integration of a function with a "rapidly varying" kernel, that is, a multiplier such as the exponential function, with real and imaginary parts each having intervals of positive function values closely adjacent to intervals of negative function values. When the amplitude function does not vary as rapidly as the kernel, the integral over a positive lobe tends to cancel the integral over the adjacent negative lobe. The cancellation is slightly less when the rapid variation is diminished, that is, when the phase is stationary. The stationary phase formula then approximates the integral over an interval around such a stationary point. The resulting Eq. (18) states that the integral over the entire interval is dominated by contributions from the neighborhoods of the stationary points.

Next, we will apply the stationary phase formula to the integrals such as those in Eq. (13). We will not always bother to rescale that Fourier representation or to introduce a dimensionless variable of integration η. We shall proceed formally in our dimensional variables, with the understanding that a complete justification of our asymptotic approximation relies on an analysis such as the one presented here. Thus, we will apply the results of this section with η replaced by k and Λ set equal to unity.

E. Asymptotic Analysis of Fourier Superposition

We will now apply the method of stationary phase of the previous section to the integral in Eq. (13). To do so, we set

$$\Phi(k) = kx - \omega(k)t \tag{18}$$

and differentiate

$$\frac{d\Phi}{dk} = x - \frac{d\omega}{dk}t; \quad \frac{d^2\Phi}{dk^2} = -\frac{d^2\omega}{dk^2}t \tag{19}$$

In accordance with Eq. (16), we set the first derivative equal to zero to determine the stationary points:

$$x = (d\omega/dk)t \tag{20}$$

The function $d\omega/dk$ was defined earlier to be the group velocity (this derivative can be positive or negative) at the given value of k. We see here that, for a given value of x and t, the stationary points are those k values for which the corresponding wave component would propagate at the group velocity $d\omega(k)/dk$ from the origin to the point x in time t. We remind the reader that the method of stationary phase provides an approximation to the integral over an interval around the stationary point. Thus, the condition of stationarity predicts that the packet of wavenumbers around the stationary value will propagate at the group velocity of the stationary value. This theme will repeat itself in higher dimensions.

We now write the asymptotic approximation of Eq. (18) to the integral of Eq. (13) as

$$u(x,t) \sim \sum_{j=1}^{n} \frac{A(k_j)}{\sqrt{2\pi\,|\omega''(k_j)|t}} \exp\{i[k_j x - \omega(k_j)t] - i(\pi/4)\mathrm{sgn}(\omega''(k_j))\} \tag{21}$$

In this equation, the summation is to be carried out over the solutions of the equation of stationarity [Eq. (20)]. We see here that each term of the sum has the structure of the fundamental waveform of Eq. (1), except that the real-valued amplitude and cosine functions have been replaced by a complex-valued amplitude and complex exponential. That is, asymptotically, the general Fourier superposition

of elementary waves behaves locally like the elementary wave, except that the phase and group velocities of the elementary waves will vary with both position and time.

This observation suggests an alternative manner in which to interpret the result of Eq. (21). Let us fix the value of k. Then we think of the packet of wavenumbers in the neighborhood of that k value as propagating at the group velocity $d\omega(k)/dk$ with amplitude and phase being given by the summand of Eq. (21) evaluated at k. For some applications, this interpretation is as useful as the actual evaluation at a given $(x,\ t)$ as defined by the summation in Eq. (21). Indeed, this interpretation provides a quantification of the definition of a wave. We see here a phase function whose crests propagate as time progresses, while the amplitude of the wave, providing the height of the crests, also varies as time progresses.

It may not be apparent why the propagation originates from the origin for this example. To understand why this is so, let us consider the wave represented by Eq. (13) at $t=0$:

$$u(x,\ 0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} A(k)\exp(ikx)\,dk \tag{22}$$

Let us rewrite this integral in terms of the dimensionless variable $\eta = k/k_0$:

$$u(x,\ 0) = \frac{k_0}{2\pi}\int_{-\infty}^{\infty} A(\eta k_0)\exp(ik_0x\eta)\,d\eta \tag{23}$$

As the product k_0x approaches infinity, the integral will approach zero under relatively mild assumptions on the amplitude A. (The Reimann–Lebesgue lemma guarantees this result if $|A(k)|$ is integrable.) Thus, we might expect that $u(x,\ 0)$ will be small for large values of k_0x and will be substantially different from zero only in some interval around the origin in which k_0x is not large. Consequently, to the order of approximation consistent with our asymptotics, the propagation of $u(x,\ t)$ initiates from the neighborhood of the origin in x. In application, the Fourier representation may well contain other terms in the phase that distribute the initiation point of different components of the wave $u(x,\ t)$ over a range of x values. For example, we might replace $A(k)$ in Eq. (13) by $A(k)\exp[i\phi_0(k)]$. We would then add derivatives of ϕ_0 to the right sides in Eq. (19). In particular, $-\phi_0'(k)$ would replace the origin as the initial value of x in Eq. (20). However, even in those cases, the propagation of the constituent elements of $u(x,\ t)$ would still be governed by the group velocity, as was the case for this simple example.

F. An Example of Dispersive Wave Propagation

We will discuss a simple example of wave propagation that will exhibit some of the features we have described in the previous section.

Let us suppose that $u(x,\ t)$ is a solution of the following initial value problem:

$$\begin{gathered}\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} + b^2u = 0, \quad t > 0, \quad -\infty < x < \infty \\ u = 0, \quad \frac{\partial u}{\partial t} = \delta(x), \quad t = 0\end{gathered} \tag{24}$$

The function $\delta(x)$ is the Dirac delta function.

We will solve the problem for u by Fourier transform. Thus, we introduce

$$\tilde{u}(k,\ t) = \int_{-\infty}^{\infty} u(x,\ t)\exp(-ikx)\,dx \tag{25}$$

By applying Fourier transform to the problem of Eq. (25), we obtain the following problem for $\tilde{u}$:

$$\begin{gathered}d^2\tilde{u}/dt^2 + (c^2k^2 + b^2)\tilde{u} = 0, \quad t > 0 \\ \tilde{u} = 0, \quad d\tilde{u}/dt = 1, \quad t = 0\end{gathered} \tag{26}$$

We leave it to the reader to verify that the solution to this initial value problem is

$$\begin{gathered}\tilde{u}(k,\ t) = \frac{\exp[i\omega(k)t] - \exp[-i\omega(k)t]}{2i\omega(k)} \\ \omega(k) = \sqrt{c^2k^2 + b^2}\end{gathered} \tag{27}$$

In this equation, we have allowed a slight abuse of notation. There are really two waves represented here: one with $\omega = \omega(k)$ and the other with $\omega = -\omega(k)$. Because the two dispersion relations define ω with only a difference in sign, we have introduced only one function $\omega(k)$.

We take the inverse Fourier transform of the solution in Eq. (28) to obtain an integral representation of the solution to the problem of Eq. (25):

$$u(x,\ t) = \frac{1}{4\pi i}\sum_{\pm}\pm\int_{-\infty}^{\infty}\frac{\exp[i\Phi_{\pm}(k,x,t)]}{\omega(k)}\,dk \tag{28}$$

where

$$\Phi_{\pm}(k,x,t) = kx \mp \omega(k)t = kx \mp \sqrt{c^2k^2 + b^2}t$$

Furthermore, the summation notation means that we add together the results for the upper and lower signs.

As a basis for comparison, it is worthwhile at this juncture to specialize the result here to the case in which there is no dispersion. That is, we consider the special case in which $b=0$ and $\omega = \pm ck$. We then find that the solution of Eq. (28) becomes

$$u(x,\,t) = \frac{1}{4\pi i c}\sum_{\pm} \pm \int_{-\infty}^{\infty} \frac{\exp[ik(x \mp ct)]}{k}\,dk$$

$$= \left\{ \begin{matrix} 1/2c, & |x| < ct \\ 0, & |x| > ct \end{matrix} \right\} = \frac{1}{2c} H(ct - |x|) \quad (29)$$

In the final expression, $H(x)$ is the Heaviside function, defined to be equal to zero for $x < 0$ and equal to one for $x > 0$. Its value at $x = 0$ is unimportant; however, if it is obtained by Fourier inversion, the value at $x = 0$ will be equal to one-half. The Fourier transform in this equation can be carried out by standard techniques of complex contour integration, or the result may be found in a standard table of Fourier transforms. We see here that the initial impulse has caused the value of $u(x,\,t)$ to "jump" from zero to the value $1/2c$ everywhere on the "characteristic" interval $(-ct,\,ct)$ and to remain there for all time. One can think of the initial data, nonzero only at the origin, as propagating to the right and left at speed c and affecting the value of $u(x,\,t)$ everywhere inside the characteristic interval.

Let us now return to the dispersive wave represented by Eq. (28). This wave is by no means as easy to analyze because of the complicated form of the integrand. We will therefore resort to our asymptotic method in an attempt to reinterpret this solution, at least asymptotically, in terms of simpler functions.

This example provides us an excellent opportunity to consider the effects of scaling to dimensionless variables, as discussed in Section I.C. Thus, we introduce the new variable of integration η, defined by

$$\eta = ck/b \quad (30)$$

As a check on dimensions, we note that c has the dimensions of length/time, whereas b must have the dimensions of 1/time for each term of the original Eq. (25) to have the same dimensions. Since k has the dimension of inverse length, η is indeed dimensionless.

In terms of η, the relevant functions of the integrand in Eq. (30) take the following form:

$$\omega(k)t = bt\sqrt{\eta^2 + 1}, \quad kx = \eta x b/c$$

$$\Phi_{\pm}(k, x, t) = bt\,\Psi_{\pm}(\eta, x, t) \quad (31)$$

$$\Psi_{\pm}(\eta, x, t) = [\eta x/ct \mp \sqrt{\eta^2 + 1}]$$

We can see in this form that the large parameter emerges naturally as bt, that is, time measured in inverse units of a characteristic frequency of the original problem. Parenthetically, we note that this is also the minimum frequency of any Fourier component of the solution of Eq. (30). Furthermore, one can check that the maximum value of the η derivative of $\Psi_{\pm}$ is $|x|/ct + 1$. [It is more difficult to show from the representation of Eq. (28), but nonetheless true, that only values of $|x| \le ct$ are of interest; otherwise the integral is identically zero.] At any distance, this bound on the derivative of $\Psi_{\pm}$ approaches unity as time increases.

The asymptotic analysis of each term in the sum in Eq. (30) proceeds as in the general case. The phase speeds [Eq. (7)] and the group speeds [Eq. (12)] for the two waves are given by

$$v_\phi = \pm\frac{\sqrt{c^2k^2 + b^2}}{k}; \quad v_g = \pm\frac{c^2 k}{\sqrt{c^2k^2 + b^2}} \quad (32)$$

We see here that the phase speeds are greater in magnitude than the speed c, while the group speeds are less in magnitude than c, for every finite value of k. Both have c as limit as $|k| \to \infty$. The magnitude of the group velocity $|v_g|$ is a monotonically increasing function of $|k|$. Thus, wave packets centered around lower wavenumbers will propagate more slowly while wave packets centered around higher wavenumbers will propagate faster. On the other hand, if one could pick out waves at a particular frequency/wave number pair, those of lower wavenumber would have crests that propagate *faster* than those of higher wavenumber. In any case, we expect, then, that the shape of the initial data function will be distorted as time progresses.

We will carry out the stationary phase analysis on the phase functions $\Phi_{\pm}$ defined in Eq. (30). Thus, following the method described in Section I.E, we calculate the first and second derivatives as in Eq. (19):

$$\frac{d\Phi_{\pm}}{dk} = x \mp v_g t; \quad \frac{d^2\Phi_{\pm}}{dk^2} = \mp\frac{c^2 b^2 t}{(c^2k^2 + b^2)^{3/2}} \quad (33)$$

We now consider the condition of stationarity [Eq. (20)] for this example:

$$x = \pm c^2 kt/\sqrt{c^2k^2 + b^2} \quad (34)$$

In many applications, it is not possible to invert this condition of stationarity to determine k as a function of x and t. In those cases, we content ourselves with a parametric solution of the form of Eq. (21) subject to the condition of Eq. (20). Indeed, in the discussion following those equations, we offered an interpretation of that representation of the solution. However, in this example it is possible to explicitly solve Eq. (34), and we now proceed to do so and thereby obtain an explicit asymptotic solution for this problem under the assumptions that bt is large.

In these equations, we see that for the upper sign (+), x and k must have the same sign at the stationary point, while for the lower sign (−), x and k must be of opposite signs. With this observation, we solve Eq. (34) for the stationary values of k, namely, $\pm k_{\text{stat}}$:

$$k_{\text{stat}} = bx/c\sqrt{c^2t^2 - x^2} \tag{35}$$

We see here that there are real solutions only for $|x| \le ct$. In the limit of equality, the stationary point moves off to infinity and the entire approximation technique breaks down. In fact, using (35) to compute the second derivative in (33), we find that

$$\Phi''_{\pm}(\pm k_{\text{stat}}) = \mp\frac{(c^2t^2 - x^2)^{3/2}}{cbt^2} \tag{36}$$

In this dimensional form, the second derivative is seen to vanish in the limit, as $x \to ct$. In such a limit, the stationary phase formula is invalid. If we had followed through on the dimensionless form, using (30), then

$$bt\frac{d^2\Psi_{\pm}}{d\eta^2} = \mp bt\left[1 - \frac{x^2}{(ct)^2}\right]^{3/2} \tag{37}$$

In this form, it is clear that our original guess at a large parameter, bt, must be tempered by the additional factor on the right. Thus, for $x = 0$, the second derivatives have magnitude bt, large enough to expect asymptotics to work by assumption. On the other hand, the method must break down near the front of propagation, where the last factor in this equation or the numerator of the previous equation is nearly equal to zero. There are more exotic asymptotic expansions that describe that region, as well, but the discussion of such techniques is beyond the scope of this article.

We now calculate the functions in the general formula of Eq. (21) for the specific example of Eq. (28) using Eq. (35). The result of that calculation is

$$u(x,\ t) \sim \frac{(c^2t^2 - x^2)^{-1/4}}{\sqrt{2\pi bc}}\cos\left(\frac{b}{c}\sqrt{c^2t^2 - x^2} - \frac{\pi}{4}\right) \tag{38}$$

This result should be compared to the exact solution,

$$u(x,\ t) = \frac{1}{2c}J_0\left[\frac{b}{c}\sqrt{c^2t^2 - x^2}\right], \quad c^2t^2 \ge x^2. \tag{39}$$

Figure 6 shows the exact solution for $t = 5$, with $c = 1$, $b = 2\pi$, and $0 \le x \le 5$; Fig. 7 shows the asymptotic solution for the same values, except that $0 \le x \le 4.98$. We see here that the character of the solution to the dispersive

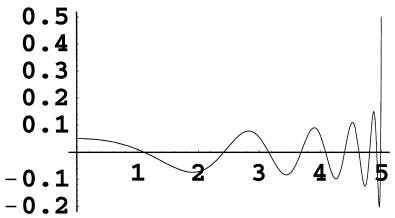


FIGURE 6 The exact solution for $t = 5$.

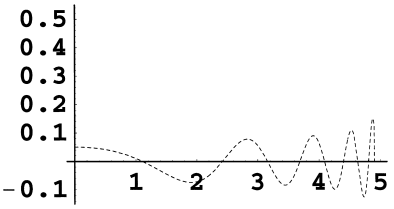


FIGURE 7 The asymptotic solution for $t = 5$.

problem is quite different from the solution, Eq. (30), to the nondispersive problem. At each fixed x, $u(x,\ t)$ now oscillates in time (as described by the cosine factor) while it decays as $1/\sqrt{t}$ as time progresses. For this problem, these are the consequences of variable propagation speed for the elements of the Fourier decomposition of the initial data.

Figure 8 shows an overlay of the two solutions. The agreement is apparent. Further, it can be seen that the wave slope increases with x. The reason is that the group velocity is a monotonic function of k. Thus, wave groups centered around large k-values propagate faster and therefore reside closer to the wave front. Larger k is related to more rapid variation and produces these larger slopes.

As noted earlier, we should expect good agreement even at $bt = \pi$. With $b = 2\pi$, that means $t = 0.5$. We show that agreement in Fig. 9. Here, again, the asymptotic solution is restricted, in this case, to an upper bound of 0.48. At least at this empirically claimed lower bound for asymptotics, some separation between the exact solution (solid curve) and the asymptotic solution (dashed curve) is visible. In fact, the difference between these two functions varies from -0.015 and $+0.015$ over the range displayed in this figure. We cannot speak of a global percentage error for these functions that pass through zero. However, the error at $x = 0$ is 4.6%; the error at $x = 0.48$ is 3.8%. Furthermore, the shift in the zero crossing between the exact and the asymptotic solution is only 0.009. In applications, the accuracy of observed data rarely matches the accuracy of the asymptotic expansion, even at this claimed lower limit of the range of validity of the asymptotic expansion. Thus,

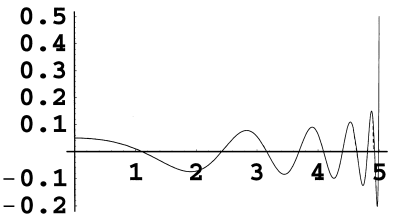


FIGURE 8 An overlay of the exact and asymptotic solutions for $t = 5$. The dashed curve is the asymptotic result, as in the previous figure.

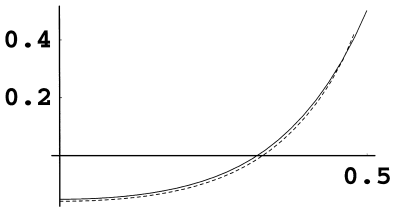


FIGURE 9 An overlay of the exact and asymptotic solutions for $bt = \pi$. The dashed curve is the asymptotic result.

the valid use of asymptotic approximations in applications is not a factor in the overall accuracy of the analysis of data.

In summary, we have seen in this example how a fairly complicated solution Eq. (28) to a dispersive wave equation [Eq. (25)] can be interpreted by asymptotic methods. In that interpretation [Eq. (38)] the distortion of the original waveform becomes more apparent and more easily recognized, especially when we compare this asymptotic solution to the exact dispersion free solution [Eq. (30)]. Furthermore, we again see the structure of a wave as described in the introduction. The equiphase points, including the crests and troughs of the wave, are determined by setting the argument of the cosine function in Eq. (38) equal to a constant. The amplitude is seen to vary both spatially and temporally. As time progresses at a fixed point x, the amplitude decays algebraically to zero, while points of constant amplitude propagate outward from the origin as time progresses.

II. WAVES IN HIGHER DIMENSIONS

We will discuss here the extension of the concepts of the previous chapter to two and three dimensions. We remark, however, that in theory there is no reason to limit our discussion to three dimensions.

We will require a notation that allows us to refer to points in two- or three-dimensional space. Thus, let us introduce the boldface symbol $\boldsymbol{x}$ to denote a point or vector in two or three dimensions. For the two-dimensional case, the coordinates of the point or the components of the vector will be $(x_1,\ x_2)$, whereas in three dimensions, $\boldsymbol{x}$ will denote the point or vector $(x_1,\ x_2,\ x_3)$. Many of the ideas we will express here will be independent of the number of dimensions.

Given two vectors $\boldsymbol{x}$ and $\boldsymbol{k}$, we will denote by $\boldsymbol{k} \cdot \boldsymbol{x}$ the *dot product* of the two vectors, defined by

$$\boldsymbol{k} \cdot \boldsymbol{x} = \sum_{j=1}^{m} k_j x_j \tag{40}$$

with m being the dimension. We will denote by x the magnitude of the vector $\boldsymbol{x}$, that is,

$$x = (\boldsymbol{x} \cdot \boldsymbol{x})^{1/2} \tag{41}$$

We will also use the notation $(\hat{\boldsymbol{x}})$ to denote the unit vector in the direction of $\boldsymbol{x}$, that is,

$$\hat{\boldsymbol{x}} = \boldsymbol{x}/x \tag{42}$$

With this notation in place, we can begin our discussion of waves in higher dimensions.

A. Plane Waves: Phase Velocity and Group Velocity

We will consider now the extension of the concepts of Section I to higher dimensions. Instead of considering the real periodic function in Eq. (1), or its alternate in which the cosine function is replaced by a sine function, we will consider here the complex exponential

$$u(\boldsymbol{x},\ t) = A \exp[i(\boldsymbol{k} \cdot \boldsymbol{x} - \omega t)] \tag{43}$$

It is to be understood that the wave we are considering is the real part of the function $u(\boldsymbol{x},\ t)$ or a real superposition of such functions.

In two dimensions, the function $u(\boldsymbol{x},\ t)$ might be thought of as the vertical displacement of a membrane or the vertical displacement of the surface of a pool of water. These are simple extensions of the concept of the vertical displacement of a string, suggested in the previous section.

For the three-dimensional case, such easily visualized wave phenomena are not available. Perhaps the easiest characterization in three dimensions might be the pressure variations or density variations of a compressible fluid, such as air. That is, one might think of sound waves. More generally, $u(\boldsymbol{x}, t)$ might represent one component of the motion of particles of an elastic medium or one component of the electric or magnetic vectors of electromagnetic propagation.

In any case, we will proceed to introduce the basic concepts of wave phenomena in higher dimensions in the context of the simple function given by Eq. (43) and its generalizations analogous to those introduced in the discussion of wave phenomena in one dimension.

Let us first consider the question of peaks of the real part of the wave of Eq. (43). These peaks are located at the positions

$$\boldsymbol{k} \cdot \boldsymbol{x} = 2n\pi + \omega t, \quad n = \ldots, -2, -1, 0, 1, 2, \ldots \tag{44}$$

For fixed t, a specific peak (fixed n) occurs everywhere on a line in two dimensions or on a plane in three dimensions. In either two or three dimensions, the wave represented by Eq. (42) is called a *plane wave*. The inclination of this plane is given by the unit normal $\hat{\boldsymbol{k}}$. The normal distance of the plane from the origin is given by $(2n\pi + \omega t)/k$, with

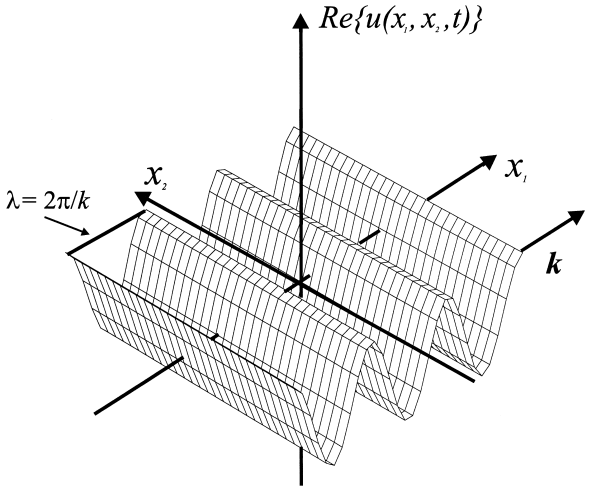


FIGURE 10 A snapshot at fixed time of a two-dimensional plane wave.

k now denoting the magnitude of $\boldsymbol{k}$. Indeed, any constant value of the phase occurs on a plane with the same features, except that the distance from the origin is determined by the specific value of the phase rather than the value $2n\pi$. All of these planes are parallel (Fig. 10).

At fixed time, the normal distance between the planes of two peak values of $u(\boldsymbol{x}, t)$ is given by

$$\lambda = 2\pi/k \tag{45}$$

Thus, we again denote by λ the wavelength of the wave represented by $u(\boldsymbol{x}, t)$. The scalar k is again called the wavenumber. The vector $\boldsymbol{k}$ is called the *wave vector*.

For fixed $\boldsymbol{x}$, the elasped time between two peaks of Re u in Eq. (43) is given by

$$T = 2\pi/\omega \tag{46}$$

As in the one-dimensional case, we call T the period of the wave and ω the frequency.

As time progresses, we can think of a plane of peak values of $\mathrm{Re}\{u(\boldsymbol{x}, t)\}$ as defined by Eq. (43) (or any plane of constant phase) as propagating normal to itself. It will propagate in the direction of $\hat{\boldsymbol{k}}$ when omega is positive or opposite to the direction of $\hat{\boldsymbol{k}}$ when ω is negative. The speed at which the plane propagates can be determined by calculating how the point on the normal through the origin propagates. That is, we set

$$\boldsymbol{x} = \hat{\boldsymbol{k}} x \ \mathrm{sgn}(\omega) \tag{47}$$

and then replace the requirement of Eq. (43) by

$$\begin{aligned} kx \ \mathrm{sgn}(\omega) &= 2n\pi + \omega t \\ n &= \ldots, -2, -1, 0, 1, 2, \ldots \\ \boldsymbol{x} &= \hat{\boldsymbol{k}} x \ \mathrm{sgn}(\omega) \end{aligned}$$

From this equation, we can see that the plane propagates normal to itself with a phase speed given by

$$v_\phi = |\omega|/k \tag{48}$$

The direction of this propagation is given by $\hat{\boldsymbol{k}}\,\mathrm{sgn}(\omega)$. Thus, we define the *phase velocity* by

$$\boldsymbol{v}_\phi = v_\phi \hat{\boldsymbol{k}}\,\mathrm{sgn}(\omega) = (\omega/k)\hat{\boldsymbol{k}} \tag{49}$$

This is the velocity with which planes of constant phase propagate.

In analogy with the one-dimensional case, let us now allow ω to be a function of $\boldsymbol{k}$, that is, $\omega = \omega(\boldsymbol{k})$. We will now consider how a wave composed of the sum of two plane waves of the form of Eq. (42) with nearby values of k might propagate. Thus, let us consider

$$\begin{aligned} u(x,\ t) = A\{&\exp[i(\boldsymbol{k}^+ \cdot \boldsymbol{x} - \omega^+ t)] \\ &+ \exp[i(\boldsymbol{k}^- \cdot \boldsymbol{x} - \omega^- t)]\} \end{aligned} \tag{50}$$

In this equation we have used $\boldsymbol{k}^\pm$ and $\omega^\pm$ as shorthand notations for

$$\begin{aligned} \boldsymbol{k}^\pm &= \bar{\boldsymbol{k}} \pm \Delta\boldsymbol{k} \\ \omega^\pm &= \bar{\omega} \pm \Delta\omega \approx \bar{\omega} \pm \nabla_k \omega(\bar{k}) \cdot \Delta\boldsymbol{k} \\ \bar{\boldsymbol{k}} &= (\boldsymbol{k}^+ + \boldsymbol{k}^-)/2, \quad \bar{\omega} = \omega(\bar{\boldsymbol{k}}) \end{aligned} \tag{51}$$

We have denoted by ∇_k the gradient of $\omega(\boldsymbol{k})$ with respect to $\boldsymbol{k}$. The dot product occurring in the approximation of $\bar{\omega}$ is the extension to two or three dimensions of the two-term Taylor expansion appearing in Eq. (10).

By using these definitions in Eq. (32) and rewriting that sum in terms of the average $\bar{\boldsymbol{k}}$ and $\Delta\boldsymbol{k}$, we obtain the following representation of the superposition of two waves:

$$u(\boldsymbol{x},\ t) = 2A\cos[\Delta\boldsymbol{k} \cdot (\boldsymbol{x} - \nabla_k \omega t)]\exp[i(\bar{\boldsymbol{k}} \cdot \boldsymbol{x} - \bar{\omega} t)] \tag{52}$$

As in the one-dimensional case, we see that the superposition of the two waves yields a wave at the average wave vector and frequency with an amplitude modulator provided by the perturbations in the average wave vector and frequency. The planes of constant phase of this modulator are of the form

$$\Delta\boldsymbol{k} \cdot (\boldsymbol{x} - \nabla_k \omega\, t) = \mathrm{const} \tag{53}$$

These planes have normal direction given by $\Delta\boldsymbol{k}$ and propagate in the direction of $\nabla_k\omega$. Indeed, the velocity of propagation is given by

$$\boldsymbol{v}_g = \nabla_k \omega \tag{54}$$

which we define to be the *group velocity*. The magnitude of this vector, $|\nabla_k\omega|$, is called the *group speed*. As in the one-dimensional case, we will see below that the group

velocity will arise in a natural way when we consider Fourier superpositions of waves in the high-frequency limit.

We remark that the phase velocity and the group velocity need not be in the same direction. Indeed, they will only be in the same direction when $\omega = \omega(k)$, that is, when omega is a function of the magnitude k rather than a function of the two or three independent components of $\boldsymbol{k}$. We list some examples of both types:

$$\begin{aligned}
&\omega = ck, && \boldsymbol{v}_\phi = c\hat{\boldsymbol{k}}, \quad \boldsymbol{v}_g = c\hat{\boldsymbol{k}} \\
&\omega = \sqrt{c^2k^2 + b^2}, && \boldsymbol{v}_\phi = \frac{\sqrt{c^2k^2+b^2}}{k}\hat{\boldsymbol{k}} \\
& && \boldsymbol{v}_g = \frac{c^2k}{\sqrt{c^2k^2+b^2}}\hat{\boldsymbol{k}} \\
&\omega = \omega_0 k_3/k, && \boldsymbol{v}_\phi = \omega_0 k_3 \big/ k^2\hat{\boldsymbol{k}} \\
& && \boldsymbol{v}_g = \omega_0\left[-k_3\big/k^2\hat{\boldsymbol{k}} + (0,\ 0,\ 1)/k\right] \\
&\omega = ck + Uk_1, && \boldsymbol{v}_\phi = [c + Uk_1/k]\hat{\boldsymbol{k}} \\
& && \boldsymbol{v}_g = c\hat{\boldsymbol{k}} + (U,\ 0\,0)
\end{aligned} \tag{55}$$

The third example here arises in the modeling of waves in a rotating fluid, and the fourth example arises in the modeling of waves in a transversely moving medium.

As in the one-dimensional case, it is the group velocity, now a vector, that governs the propagation of energy over a distance.

B. Fourier Superposition

We now consider waves that are the Fourier superposition of plane waves of the type introduced above. Thus, let us set

$$u(\boldsymbol{x},\ t) = \frac{1}{(2\pi)^m}\int_{-\infty}^{\infty} A(\boldsymbol{k})\exp\{i\ [\boldsymbol{k}\cdot\boldsymbol{x} - \omega(\boldsymbol{k})t]\}\,d^m k \tag{56}$$

In this equation, the domain of integration is understood to be from $-\infty$ to ∞ in all m independent k variables. For our purposes, m will be restricted to 2 or 3.

Such Fourier superpositions can be used to reconstruct a broad class of waves. Below, we describe three quite different types of waves and their corresponding Fourier transforms, $A(\boldsymbol{k})$, along with the necessary dispersion relation. That is, we will provide the amplitudes of the integrand in (57), as well as the attendant function, $\omega(\boldsymbol{k})$, needed to complete the integrand in that equation. In all examples, $m = 3$.

The first example is a periodic plane wave in three dimensions

$$u(\boldsymbol{x}, t) = \cos[\boldsymbol{k}_0\cdot\boldsymbol{x} - ck_0 t]$$

for which

$$\begin{aligned}
A(\boldsymbol{k}) &= A_+(\boldsymbol{k}) + A_-(\boldsymbol{k}) \\
A_\pm(\boldsymbol{k}) &= 4\pi^3\delta(k_1 \mp k_{10})\,\delta(k_2 \mp k_{20})\cdot\delta(k_3 \mp k_{30}) \\
\omega &= \omega_\pm(\boldsymbol{k}) = \pm ck_0
\end{aligned} \tag{57}$$

Here, upper signs in the last two lines go together, as do the lower signs.

The second example is the Green's function for the wave equation in a homogeneous medium:

$$u(\boldsymbol{x},\ t) = \frac{\delta(t - r/c)}{4\pi r}$$
$$r = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

for which

$$\begin{aligned}
A(\boldsymbol{k}) &= A_+(\boldsymbol{k}) + A_-(\boldsymbol{k}), \\
A_\pm(\boldsymbol{k}) &= \pm\frac{ic}{2k}, \quad \omega_\pm(\boldsymbol{k}) = \pm ck.
\end{aligned} \tag{58}$$

It should be noted that the singularity, $1/k$, in these amplitudes is actually quite mild in three dimensions, owing to the fact that the differential volume element written in spherical polar coordinates is $k^2\sin\theta\,dk\,d\theta\,d\phi$. The multiplication by k^2 in the inverse transform assures that the volume integral will not be singular at $k = 0$.

Note also that if this representation is derived as the solution of a *causal* problem, that is, one for which $u = 0$ for $t < 0$, then it should only be used for $t > 0$. If not, it will actually yield a second wave, $\delta(t + r/c)/4\pi r$, propagating backwards in time! Use of a causal inverse Fourier transform in time will ensure that this wave does not arise. Discussion of causal Fourier transforms is beyond the scope of this article.

The last example is the distributional plane wave:

$$u(\boldsymbol{x},\ t) = \delta(x_1 - ct)$$

for which

$$\begin{aligned}
A(k) &= (2\pi)^2\delta(k_2)\delta(k_3) \\
\omega(\boldsymbol{k}) &= ck_1
\end{aligned} \tag{59}$$

C. Multidimensional Stationary Phase Formula

In Eq. (57) we cannot as easily write down a closed-form recognizable function representing the wave propagating in space and time. In this case, we again resort to the method of stationary phase, this time *multidimensional stationary phase*, to approximate the multifold wave form in terms of more familiar plane waves of the form of Eq. (43) for arbitrary $A(\boldsymbol{k})$. First, we present the multidimensional stationary phase formula.

Let us suppose that the integral I is defined by

$$I(\Lambda) = \int f(\boldsymbol{\eta}) \exp[i\Lambda\Phi(\boldsymbol{\eta})]\, d^m\eta \tag{60}$$

In this equation, the single integral sign is understood to represent an m-fold integral over the m variables $\eta_1, \eta_2, \ldots, \eta_m$. We are interested in an approximation of the integral for large values of Λ.

As in one dimension, the integral is dominated by contributions from the neighborhoods of certain critical points, $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n$, called *stationary points*, where

$$\partial\Phi(\boldsymbol{\eta})/\partial\eta_p = 0, \qquad p = 1, 2, \ldots, m$$
$$\boldsymbol{\eta} = \boldsymbol{\eta}_j, \qquad j = 1, 2, \ldots, n$$

This is the generalization of the condition of Eq. (15).

A stationary point are called simple when the *Hessian matrix*, the matrix of second derivatives, has a nonzero determinant at that point. That is,

$$\det\left[\Phi_{pq}(\boldsymbol{\eta}_j)\right] \neq 0; \quad \Phi_{pq}(\boldsymbol{\eta}) = \partial^2\Phi(\boldsymbol{\eta}_j)\big/\partial\eta_p\,\partial\eta_q$$
$$p, q = 1, 2, \ldots, m, \quad j = 1, 2 \ldots, n \tag{61}$$

The integral I is then approximated by

$$I \sim \sum_{j=1}^{n} \left(\frac{2\pi}{|\Lambda|}\right)^{m/2} \frac{f(\boldsymbol{\eta}_j)}{\sqrt{|\det[\Phi_{pq}(\boldsymbol{\eta}_j)]|}} \exp[i\Lambda\Phi(\boldsymbol{\eta}_j)$$
$$+ i(\pi/4)\mathrm{sgn}(\Lambda)\mathrm{Sgn}(\Phi_{pq}(\boldsymbol{\eta}_j))] \tag{62}$$

In this equation, $\mathrm{Sgn}(\Phi_{pq})$ denotes the signature of the matrix $[\Phi_{pq}]$. The signature of a matrix is the number of positive eigenvalues minus the number of negative eigenvalues of the matrix. This result is the multidimensional stationary phase formula.

The qualitative description of the method of stationary phase is completely analogous to the discussion of the one-dimensional case. Each term in the sum in Eq. (63) is an approximation to the integral in a small domain around the stationary point.

As in the one-dimensional case discussed in Section I, a dimensionless large parameter Λ can be identified for integrals of the type in Eq. (57) by recasting that integral in dimensional variables in terms of dimensionless variables. However, we will proceed formally to use this approximation in the dimensional integral of Eq. (57) with the formal large parameter equal to unity. As we demonstrated in Section I, this will produce an asymptotic approximation valid for large time measured in units of a characteristic time of the integral or large distance measured in a characteristic distance of the integral.

D. Asymptotic Analysis of Fourier Superposition

We will now apply the multidimensional stationary phase formula of Eq. (63) to the integral of Eq. (57). To do so, we introduce the phase function

$$\Phi(\boldsymbol{k}) = \boldsymbol{k}\cdot\boldsymbol{x} - \omega(\boldsymbol{k})t \tag{63}$$

In order to use this method, both the first and second derivatives of this phase function are needed. Those derivatives are

$$\frac{\partial\Phi(\boldsymbol{k})}{\partial k_p} = x_p - \frac{\partial\omega(\boldsymbol{k})}{\partial k_p}t$$
$$\frac{\partial^2\Phi(\boldsymbol{k})}{\partial k_p\partial k_q} = -\frac{\partial^2\omega(\boldsymbol{k})}{\partial k_p\partial k_q}t = -\omega_{pq}(\boldsymbol{k})t \tag{64}$$
$$p, q = 1, 2, \quad \text{or} \quad 1, 2, 3$$

The stationary points are determined by setting the first derivatives of Φ equal to zero. We write that result in the vector form,

$$\boldsymbol{x} = \nabla_k\omega(\boldsymbol{k})t \tag{65}$$

The vector on the right side, $\nabla_k\omega(\boldsymbol{k})$, can be recognized as the *group velocity vector* introduced earlier. For a particular choice of $(\boldsymbol{x}, t)$, the stationary points in $\boldsymbol{k}$ are those points for which the group velocity is the velocity of propagation from the origin to $\boldsymbol{x}$ in the time t. We remark that with more structure in $A(\boldsymbol{k})$ (for example, some phase dependence), we could create examples in which the propagation is not from the origin but from other points in space. In any case, the velocity of propagation picked out by the condition of stationarity would remain the group velocity. Again, as in one dimension, the contribution from each stationary point, that is, each solution of Eq. (65), approximates the integral in a local domain around the stationary point. Thus, each such contribution represents the propagation of a packet of wave vectors in a neighborhood of the particular wave vector satisfying Eq. (65).

The asymptotic approximation to Eq. (57) in the form of Eq. (63) is

$$u(\boldsymbol{x}, t) \sim \sum_{j=1}^{n} \frac{1}{(2\pi t)^{3/2}} \frac{A(\boldsymbol{k}_j)}{\sqrt{|\det[\omega_{pq}(\boldsymbol{k}_j)]|}}$$
$$\times \exp\{i[\boldsymbol{k}_j\cdot\boldsymbol{x} - \omega(\boldsymbol{k}_j)t]$$
$$- i(\pi/4)\mathrm{Sgn}(\omega_{pq}(\boldsymbol{k}_j))\} \tag{66}$$
$$\boldsymbol{x} = \nabla_k\omega(\boldsymbol{k}_j)t$$

We see here that asymptotically each term of this general wave form behaves locally as a plane wave propagating at a group velocity that, in general, will vary from point to point in space. This is an essential feature of high-frequency propagation of waves. Thus, the propagation of

plane waves takes on an added significance as the local propagation of more complex wave structures.

As in the one-dimensional case, we see in the structure of this representation a wave with recognizable crests—which are the phase surfaces of the exponential—and slowly varying amplitude.

The propagation paths along which the solution propagates [Eq. (65)] turn out to be the rays of geometrical optics, a high-frequency technique based on the WKBJ method for ordinary differential equations. For continuous gradient functions $\nabla_k \omega(\boldsymbol{k})$, the rays for a packet of nearby $\boldsymbol{k}$ values will remain near to one another and will fill out a cone (not necessarily of circular cross section) as time progresses.

This observation leads to the interpretation of the solution representation as an example of energy conservation. Returning to the representation of Eq. (57) and setting $t = 0$, we see that $A(\boldsymbol{k})$ can be interpreted as the spectral density of the initial data. We then think of the square of this quantity, $|A(\boldsymbol{k})|^2$, as being the spectral density of the energy in the k domain or $|A(\boldsymbol{k})^2|\Delta V_k$ as the energy in the packet of k values in the volume element ΔV_k around $\boldsymbol{k}$.

Let us define $|A(\boldsymbol{k},\ t)|$ to be the amplitude of the wave $u(\boldsymbol{x},\ t)$ at fixed $\boldsymbol{k}$ as time progresses. In this expression of the amplitude, we define the $\boldsymbol{x}$-coordinate associated with $\boldsymbol{k}$ by the ray equation (65). Thus, $|A(\boldsymbol{k},\ 0)|$ is just the spectral density $|A(\boldsymbol{k})|$. In an energy-conserving system, we expect that as the wave propagates, $|A(\boldsymbol{k},\ t)|^2 \Delta V_k(\boldsymbol{k},\ t)$ will be preserved (that is, remain constant) while the volume element varies in accordance with the ray equation (65). The product $t^3|\det(\omega_{pq}(\mathbf{k}))|$ is the Jacobian of transformation via rays and is proportional to this volume element. Thus, for the energy to be preserved in a packet of k values, the energy density $|A(\boldsymbol{k},\ t)|^2$ must vary inversely with this Jacobian, and the amplitude $|A(\boldsymbol{k},\ t)|$ of the wave must therefore vary inversely with the square root of this Jacobian. This provides a physical interpretation of the division by the square root of the Hessian matrix in the asymptotic expressions of the summand in Eq. (66a) and our interpretation of the solution formula as a manifestation of conservation of energy. It is also consistent with our earlier claim that energy propagates at the group velocity.

E. Plane Waves: Reflection and Refraction

Fundamental to the set of concepts of how plane waves propagate is the interaction of such waves with a planar boundary across which some property of the medium of propagation [equivalently, some coefficient(s) of the modeling equation(s)] changes. We will describe this phenomenon in the context of a specific example and then discuss generalizations of the basic result.

Let us suppose that we are considering plane waves that are solutions of the wave equation

$$c^2\left[\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2}\right] - \frac{\partial^2 u}{\partial t^2} = 0$$

$$c = \begin{cases} c_-,\ x_1 < 0 \\ c_+,\ x_1 > 0 \end{cases} \tag{67}$$

We wish to consider the interaction of a plane wave at a fixed frequency, incident on the interface at $x_1 = 0$ from the left; that is, from the medium in which $c = c_-$. Thus, we anticipate an incident wave, which we will denote by u_I of the form

$$u_I(\boldsymbol{x}, t) = A_I \exp\{i[(\boldsymbol{k}_I \cdot \boldsymbol{x} - \omega(\boldsymbol{k}_I)t]\} \tag{68}$$

In this equation, we must choose $\omega(\boldsymbol{k})$ so that the plane wave satisfies the governing equation (67). Thus,

$$\omega^2 = c_-^2 k_I^2, \quad \omega = \pm c_- k_I \tag{69}$$

Of the two choices, we will set $\omega = ck_I$. This was the first example of a dispersion relation in Eq. (55). With this choice, both the phase velocity and the group velocity have the same direction as $\boldsymbol{k}_I$; for the opposite choice, the two velocities would be directed opposite to $\boldsymbol{k}_I$. Thus, so that our plane wave is propagating from $x_1 < 0$ toward $x_1 = 0$, $\boldsymbol{k}_I$ must make an acute angle with the x_1 axis. That is,

$$k_{1I} > 0, \quad \omega = ck_I \tag{70}$$

We will conjecture that the total solution in $x_1 < 0$ is made up of the incident wave and another wave called the *reflected wave* (u_R). Furthermore, we will assume that another plane wave is transmitted (u_T) through the interface. Thus, we conjecture a total solution of the form

$$u(\boldsymbol{x}, t) = \begin{cases} u_I(\boldsymbol{x}, t) + u_R(\boldsymbol{x}, t),\ x_1 < 0 \\ u_T(\boldsymbol{x}, t), \qquad\qquad x_1 > 0 \end{cases}$$

$$u_R(\boldsymbol{x}, t) = A_R \exp\{i\,[\boldsymbol{k}_R \cdot \boldsymbol{x} - \omega_R t]\} \tag{71}$$

$$u_T(\boldsymbol{x}, t) = A_T \exp\{i\,[\boldsymbol{k}_T \cdot \boldsymbol{x} - \omega_T t]\}$$

Our objective now is to express ω_R, ω_T, $\boldsymbol{k}_R$, $\boldsymbol{k}_T$, A_R, and A_T in terms of $\boldsymbol{k}_I$ and A_I. That is, we seek to express the frequencies, the directions of propagation, and the amplitudes of the reflected and transmitted waves in terms of the same parameters for the incident wave and conditions imposed on the model as to how these waves are to interact at the boundary.

A typical requirement of such interactions is that the solution be continuous across the interface. That is,

$$\begin{aligned} &A_I \exp[i(k_{2I}x_2 + k_{3I}x_3 - c_- k_I t)] \\ &\quad + A_R \exp[i(k_{2R}x_2 + k_{3R}x_3 - \omega_R t)] \\ &\quad = A_R \exp[i(k_{2T}x_2 + k_{3T}x_3 - \omega_T t)] \end{aligned} \tag{72}$$

We take the Fourier transform of this equation with respect to t, that is, we multiply by $\exp(i\mu t)$ and integrate from $-\infty$ to ∞ with respect to t, and we find that all frequencies must agree. This follows from the fact that the first integral is proportional to $\delta(\mu - ck_I)$, whereas the second is proportional to $\delta(\mu - \omega_R)$ and the third is proportional to $\delta(\mu - \omega_T)$. Since each of these Dirac delta functions is nonzero only where its argument is zero, they could not agree unless the frequencies were the same. Thus,

$$\omega_R = c_- k_R = \omega_T = c_+ k_T = c_- k_I \tag{73}$$

By a completely analogous argument applied to the spatial transforms, we find also that

$$k_{2R} = k_{2T} = k_{2I} \qquad \text{and} \qquad k_{3R} = k_{3T} = k_{3I} \tag{74}$$

These equations state that the projections of the three wave vectors on the planar interface must agree. The previous equation, in addition to equating the frequencies, states that the magnitudes of the reflected wave vector must equal the magnitude of the incident wave vector, while the magnitude of the transmitted wave vector must equal these two up to a scale factor.

Let us first focus our attention on $\boldsymbol{k}_R$, the reflected wave vector. From Eqs. (73) and (74) it follows that $k_{1R} = \pm k_{1I}$. If these two components had the same sign, then $\boldsymbol{k}_R$ would equal $\boldsymbol{k}_I$ and the reflected wave would also be directed toward the interface. On physical grounds, we reject this; we expect u_R to be a wave directed away from the interface. The mathematical basis for rejecting this case is equally strong. Were we to continue, we would find that A_R would be the negative of A_I and A_T would be zero. That is, a total solution that is identically zero would result. This is not the solution of interest. Thus, whether on mathematical grounds or physical grounds, we set

$$k_{1R} = -k_{1I} \tag{75}$$

We see then that the incident and reflected wave vectors differ only in the sign of the normal component. Thus, these two vectors must make equal angles with the normal vector to the interface. This is Snell's law of reflection.

Let us now consider the parameters for the transmitted wave. We denote by K_I and K_T, respectively, the magnitudes of the transverse components of the wave vectors $\boldsymbol{k}_I$ and $\boldsymbol{k}_T$:

$$K_I = \sqrt{k_{2I}^2 + k_{3I}^2}, \quad K_T = \sqrt{k_{2T}^2 + k_{3T}^2} \tag{76}$$

From Eq. (74), we see that these two magnitudes are equal. Furthermore, dividing this equality by the last part of Eq. (73) yields

$$\frac{K_I}{c_- k_I} = \frac{K_T}{c_+ k_T} \tag{77}$$

This is Snell's law of refraction, and the transmitted wave is, in fact, the refracted wave. The law is more often expressed in terms of the angles of incidence and refraction, these being the angles that the wave vectors make with the normal to the interface. If we denote those angles by I and R, respectively, then

$$\sin I = K_I / k_I, \quad \sin R = K_R / k_R \tag{78}$$

Thus, we conclude from Eqs. (77) and (78) that

$$\sin R / \sin I = c_+ / c_- \tag{79}$$

This is Snell's law of refraction in more familiar form. In order that R be a real angle, we must require that $\sin R$ be less than or equal to unity. Equivalently, we require that $(c_+/c_-)\sin I \le 1$. When this criterion is violated (only possible for $c_+ > c_-$), we do not have a wave of the form of Eq. (72) propagating in the second medium.

We now determine k_{1T}. From Eq. (73), we can see that

$$k_T^2 = k_{1T}^2 + k_{2T}^2 + k_{3T}^2 = c_-^2 k_I^2 / c_+^2 \tag{80}$$

We know k_{2T} and k_{3T} from Eq. (74). Thus, we can determine k_{1T} within a sign. We require that u_T be a wave propagating away from the interface. Thus, k_{1T} must be positive, and the solution for k_{1T} is

$$\begin{aligned} k_{1T} &= \sqrt{k_I^2 c_-^2 / c_+^2 - k_{2I}^2 - k_{3I}^2} \\ &= k_I \sqrt{c_-^2 / c_+^2 - \sin^2 I} \end{aligned} \tag{81}$$

Our assumption that the angle of refraction be real assures us that k_{1T} is real. We can now see that when this criterion is violated, k_{1T} is imaginary and an *attenuated or evanescent wave* propagates in the second medium.

In summary, determination of the direction of propagation of the reflected and refracted wave rests totally on the matching of the phases at the interface. Thus, even under conditions that require some multiple of $u(\boldsymbol{x}, t)$ on both sides of the interface to be equal, the same conclusion would be reached. Furthermore, we can state this result in more general terms. First, Eq. (73) tells us that the frequencies of all of the waves must agree at the interface. Since the frequency is related to the wave vectors through the disperion relation, we obtain one equation relating the wave vector $\boldsymbol{k}_R$ to $\boldsymbol{k}_I$ and another relating $\boldsymbol{k}_T$ to $\boldsymbol{k}_I$. In general, these equations are nonlinear. Equation (74) may be viewed as prescribing that the projections of all of the wave vectors on the interface (i.e., the transverse part of the wave vectors) must agree. This provides another pair of equations for the components of $\boldsymbol{k}_R$ and another pair of equations for the determination of $\boldsymbol{k}_T$. Indeed, this determines the transverse components of the wave vectors, and only the normal component remains to be determined. It is in this normal component that all of the change from $\boldsymbol{k}_I$ in the structure of the wave vectors

can occur. Finally, we observe that for our high frequency approximation [Eq. (66)] to the general Fourier superposition, the same result obtains in a pointwise manner at the interface. These features are common to all linear wave phenomena.

To determine the amplitudes A_R and A_T in Eq. (72), we need a second relationship between the solutions on the two sides of the interface. We will impose the condition that the normal derivatives of the fields be equal at the interface. For our specific example of an interface at $x_1 = 0$, the normal derivative is the x_1 derivative. We will differentiate the two representations of $u(\boldsymbol{x},\ t)$ in Eq. (72) and then set x_1 equal to zero. We exploit what we already know about the wave vectors and frequency to simplify this expression. We also use Eq. (72) with the same simplifications. This leads to a pair of equations in the two unknowns A_R and A_T:

$$\begin{aligned} A_I + A_R &= A_T \\ k_{1I} A_I - k_{1I} A_R &= k_{1I} \sqrt{c_-^2/c_+^2 - \sin^2 I}\, A_T \end{aligned} \tag{82}$$

The solution of this pair of equations is

$$A_R = R A_I, \quad A_I = T A_I \tag{83}$$

where R and T are, respectively, the *reflection coefficient* and *transmission coefficient*, which relate the amplitudes A_R and A_T to A_I. They are given by

$$\begin{aligned} R &= \frac{1 - \sqrt{c_-^2/c_+^2 - \sin^2 I}}{1 + \sqrt{c_-^2/c_+^2 - \sin^2 I}} \\ T &= \frac{2}{1 - \sqrt{c_-^2/c_+^2 - \sin^2 I}} \end{aligned} \tag{84}$$

The value of $\sin I$ in terms of $\boldsymbol{k}_I$ is given by Eq. (78). At normal incidence, that is, when the angle I is zero, $\sin I = 0$, these coefficients reduce to

$$\begin{aligned} R &= \frac{c_+ - c_-}{c_+ - c_-} \\ T &= \frac{2c_+}{c_+ - c_-} \end{aligned} \tag{85}$$

SEE ALSO THE FOLLOWING ARTICLES

ACOUSTICS, LINEAR • ATMOSPHERIC TURBULENCE • ELECTROMAGNETICS • FOURIER SERIES • GREEN'S FUNCTIONS • PHYSICAL OCEANOGRAPHY, OCEANIC ADJUSTMENT • PLANETARY WAVES • RADIO PROPAGATION • SEISMOLOGY, THEORETICAL • TIME AND FREQUENCY

BIBLIOGRAPHY

Aki, K., and Richards, P. (1980). "Quantitative Seismology: Theory and Methods," Vols. 1 and 2, Freeman, New York.

Bleistein, N. (1984). "Mathematical Methods for Wave Phenomena," Academic Press, New York.

Bleistein, N., and Handelsman, R. A. (1986). "Asymptotic Expansions of Integrals," Dover Publications Inc., New York.

Bleistein, N., Cohen, J. K., and Stockwell, J. W., Jr. (2000). "Mathematics of Multidimensional Imaging, Migration and Inversion," Springer-Verlag, New York.

Brekhovskikh, L. M., and Godin, O. A. (1998). "Acoustics of Layered Media I : Plane and Quasi-Plane Waves," Springer-Verlag, New York.

Brillouin, L. (1960). "Wave Propagation and Group Velocity," Academic Press, New York.

Brillouin, L. (1953). "Wave Propagation in Periodic Structures," Dover, New York.

Erdélyi, A. (1954). "Asymptotic Expansions of Integrals," Dover Publications Inc., New York.

Ewing, W. M., Jardetzky, W. S., and Press, F. (1957). "Elastic Waves in Layered Media," McGraw-Hill, New York.

Felsen, L. B., and Marcuvitz, N. (1973). "Radiation and Scattering of Waves," Prentice Hall, Englewood Cliffs, NJ.

Goodman, J. W. (1968). "Introduction to Fourier Optics," McGraw-Hill, New York.

Jackson, J. D. (1998). "Classical Electrodynamics," 3rd ed., John Wiley & Sons, New York.

Pekeris, C. L. (1963). "Theory of propagation of explosive sound in shallow water," *in* "Propagation of Sound in the Ocean," *Geol. Soc. Am. Memoir 27*.

Sommerfeld, A. (1964). "Optics, Lectures on Theoretical Physics," Vol. 4, Academic Press, New York.

Stoker, J. J. (1957). "Water Waves," Wiley (Interscience), New York.

Titchmarsh, E. C. (1948). "Introduction to the Theory of Fourier Integrals," Clarendon, Oxford.

Whitham, G. B. (1974). "Linear and Nonlinear Waves," Wiley, New York.