The New Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 17

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

Encyclopædia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974

Chicago
Auckland/Geneva/London/Manila/Paris/Rome
Seoul/Sydney/Tokyo/Toronto

First Edition 1768-1771
Second Edition 1777-1784
Third Edition 1788-1797
Supplement 1801
Fourth Edition 1801-1809
Fifth Edition 1815
Sixth Edition 1820-1823
Supplement 1815-1824
Seventh Edition 1830-1842
Eighth Edition 1852-1860
Ninth Edition 1875-1889
Tenth Edition 1902-1903

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984
By Encyclopædia Britannica, Inc.

© 1984
By Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Catalog Card Number: 82-84048
International Standard Book Number: 0-85229-413-1

Sonar

Sonar is the name of a technique for detecting the presence of objects underwater by acoustical echo. Having been developed during World War I, it antedates the better known radar, which uses electromagnetic echo. "Active sonar" employs an apparatus for radiating acoustical energy to bounce off underwater objects; "passive sonar" consists merely of the passive reception of acoustical energy generated by another source.

DEVELOPMENT OF THE TECHNIQUE

Interest in a means of detecting underwater objects was originally aroused by the problem of icebergs, dramatized by the sinking of the "Titanic" in 1912. The first proposal was put forward by a British meteorologist, L.F. Richardson, and the first successful application used in iceberg detection was made by the American radio pioneer, R.A. Fessenden. Development was stimulated by the outbreak of World War I and the impact of submarine warfare. A French physicist, Paul Langevin, played the leading role in research in which first British and then American scientists joined. A passive system of submarine detection, operational by 1916, employed a hydrophone (underwater microphone) and amplifier to pick up the noise emitted by submarine engines. By 1918 scientists had developed an active system in which a pulse of sound was transmitted and its rebounding echo used to detect a submarine even when its engines were shut down. The original term, "asdics," is said to have been derived from 'Anti-Submarine Division-ics,' although other explanations have been given. The name was long retained in the United Kingdom; the term sonar, from sound navigation and ranging, a United States acronym from World War II, is now used widely.

In the years between the wars, British and United States researchers refined techniques to such a point that the Allies enjoyed a substantial advantage over Germany in World War II underwater detection. They developed two types of beams, one vertical to bounce off the sea floor for depth determination, the other horizontal and capable of detecting underwater objects and obstacles near the surface. The depth detector, or echo sounder, was in widespread navigational use in the 1930s. A major American wartime contribution was the development of a system that could rapidly scan with a narrow beam either a sector or all around, without mechanical motion of the acoustic transmitters or receivers, making possible swift and methodical search for submarines.

In the years since 1945 a great deal of work has been done in naval acoustics by the United States, the United Kingdom, and the Soviet Union, and other maritime nations, but virtually all of it is secret and so has had little influence on peaceful applications. The most important of these is the use of echo sounders to detect shoals of fish, a potential discovered in the 1930s. By 1950 specialized equipment was being installed on fishing vessels, with notable success.

The echo sounder

In the early 1970s the simple echo sounder remained the basic sonar device used by fishing fleets. A type of sonar has been developed with a horizontal scan that locates fish at a distance of one kilometre (about 1,100 yards), greatly facilitating purse-seining (large net) operations by trawlers. The beam's direction can be changed by mechanical rotation of the underwater transmitting and receiving equipment, permitting a thorough search of a fair-sized area. The technique is slow, owing to the slow rate at which sound travels through water (1.5 kilometres per second).

Another problem in fish sonar is that there is as yet no technique for detecting fish close to the seabed in front of the ship, as the bottom-trawling method of fishing used by the British and other fishermen requires. There are, however, systems that accurately indicate the depth of a net towed behind the trawler, so that adjustments can be made to take advantage of any fish shoals detected on the main sonar.

Known military applications in the 1970s include the detection and location of submarines, control of antisubmarine weapons, sonar-equipped homing torpedoes, and mine hunting.

Detection ranges for civil and military sonar systems vary from 100 metres to 10 kilometres. Wave lengths for the acoustic signal range between 0.5 centimetres and 30 centimetres, corresponding to frequencies of approximately 300 kilohertz and 5 kilohertz.

BASIC PHYSICAL PRINCIPLES

Principles of a simple sonar system. As with most radar systems, sonar systems generally use the transmission and reflection of a pulse of energy as their basis of operation. The arrangement in Figure 1 is typical. Individ-

From B.K. Gazey and D.G. Tucker, *Applied Underwater Acoustics* (1966); Pergamon Press; reprinted by permission

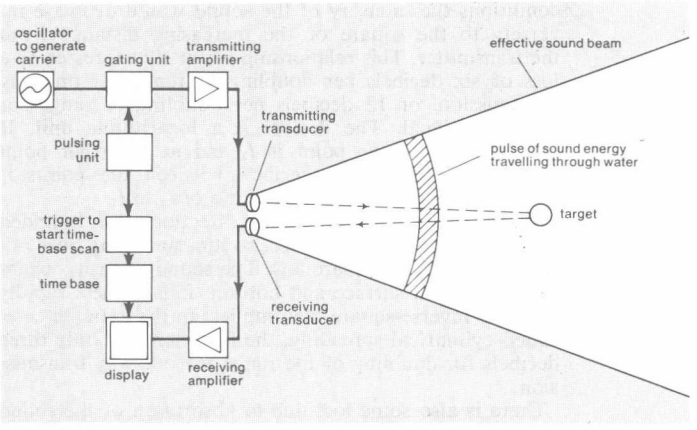


Figure 1: Schematic arrangement of a typical pulsed sonar system.

ual systems vary, with some, for example, using a single transducer both for generating the acoustic waves in the water and for detecting the reflected waves. The pulses they send out are always bursts of the transmitted, or carrier, frequency, although the term sonar may also include marine seismic work, using the sound pulses generated by a small explosion.

Display. The distance from the transducers to a reflecting target is indicated by the time elapsing between transmission of the pulse and the reception of an echo. The display, usually a pattern on a moving strip of paper, always includes a time base, the traverse of which is initiated by the transmitted pulse. For simple systems, the commonest display is a chemical recorder which records the received information on sensitized paper by means of a recording stylus drawn across paper. Since the echo actuates the marking mechanism, the position of the mark across the traverse indicates the range of the target. The

paper is moved slowly in a direction perpendicular to the traverse of the marking stylus to make successive traverses side by side. If the range of the target in relation to the transducers does not vary, the line produced by the echo is parallel to the direction of motion of the paper. If the range changes (due to the motion of the target or ship), the line slopes with respect to the paper motion.

Resolution. The resolution of the system in range, that is, the fineness of detail it can show in a direction outward from the transmitter, improves as the transmitted pulse is made shorter. The carrier frequency may have to be increased to permit this shortening. Greater bandwidth is then required in the receiving equipment. Improved angular resolution, that is, the ability to distinguish two targets at slightly different angles, is generally obtained by making the beams narrow (more directional).

Doppler sonar. Another sonar method, "Doppler detection," relies upon the relative speed of the target and the observing station to provide an indication of target speed. It employs the Doppler effect, in which an apparent change in frequency occurs when the observed and observer are in motion relative to one another. The classic example is the apparent change in the pitch of a train whistle as the locomotive approaches and passes an observer. Detection by this means requires a long pulse to give an aural recognition of tone. The received signal is mixed with a local signal of slightly different frequency, the frequencies being so arranged that mixing the two produces an audible signal. A shift of pitch between the transmitted and received signals can easily be detected by a trained ear; this method gives not only very sensitive detection but also valuable information regarding the speed of the target.

Doppler detection can be used on fast-moving ocean fish, but most fish move too slowly for the method to be of value for ordinary fishing.

The propagation of acoustic waves in water. If water did not introduce any losses, if it extended indefinitely in all directions, and if it were uniform in all respects, then the spreading of sound would be spherical. Under these conditions the intensity of the sound would decrease inversely to the square of the increasing distance from the transmitter. This relationship is usually expressed as a loss of six decibels per doubling of range for one-way transmission, or 12 decibels per doubling of range for the echo signal. The decibel is a logarithmic unit. If the intensity at one point is I_1 and at a farther point is I_2, then the loss in decibels between the points is ten times the logarithm of the ratio of I_1 to I_2.

Sound intensity in the sea

But the sea is not infinite in all directions. It is bounded by the surface and the seabed, so that spreading does not follow the inverse-square law. The sound intensity, channelled between surface and bottom, falls off less rapidly than the inverse-square law implies. In the extreme case, called cylindrical spreading, the loss increases only three decibels for doubling of the range for one-way transmission.

There is also some loss due to absorption of the sound energy by the water, which converts it into heat. This loss is variable; it is small at low frequencies but rises very rapidly with an increase in frequency. In freshwater, the loss expressed in decibels per kilometre is proportional to the square of the frequency and is about three at 100 kilohertz. In seawater, there is an additional loss at frequencies below one megahertz due to the dissolved salts. Below 100 kilohertz, this additional loss is approximately constant at around 15 decibels per kilometre.

Further propagation effects of importance result from variations of the velocity of sound in seawater. Under normal conditions the velocity is about 1,500 metres per second. The main causes of varying velocity are, in order of their importance: temperature, pressure due to depth, and salinity. The magnitudes of the effects cannot be expressed by any fundamental equation, though several empirical, or rule-of-thumb, equations have been proposed. Increased salinity, temperature, and depth all increase velocity over all normal ranges.

Perhaps the most serious effect of varying velocity is the refraction (bending) of the sound beam, which may cause a beam that is normally horizontal to be deflected to the sea bottom, where it will be reflected upward, then down again, and so on. Along any straight line through the transducer, therefore, there are intervals of range where detection is impossible and others where it is possible. This and similar effects are most serious in low-frequency systems since they have the greatest nominal range.

Effects of varying velocity

Another effect of varying velocity is the general scattering of the sound beam as it passes through turbulent water, leading to rapid fluctuations of signal strength.

INHERENT LIMITATIONS OF SONAR

Noise in water. Acoustic noise in water sets the ultimate limit to the range of detection because it eventually obscures the wanted signal. Noise is defined as power received that is not part of the desired signal and is not produced by the transmitter. Sources of noise can be inherent, can be caused by natural phenomena, or can be produced by man or animals.

Inherent noise. Inherent noise results from motion of molecules, caused by heat. The colder the water the less inherent noise is received. Basically, noise intensity from this source is independent of frequency although, like all noise, it is dependent on the bandwidth of the system.

Sea-state noise. The most important noise produced by natural phenomena is that resulting from wave action. This is called "sea-state noise." Its magnitude is dependent on the height of the waves. Essentially a low-frequency noise, its power diminishes rapidly as the frequency rises, becoming negligible in comparison with thermal noise around 50 to 150 kilohertz.

Animal and man-made noises. Noise-producing aquatic animals include several species of fish and shrimp, encountered frequently in warm waters. Man-made noise, particularly that generated by the ship carrying the sonar equipment, can be more serious. Although difficult to calculate in advance, it poses the main limitation to sonar performance.

Reverberation. When the sonar system must detect small targets, a random background due to "reverberation" can make detection difficult. Reverberation is the sum of all the numerous small echoes produced by reflections from sand and stone particles on the sea bottom, minute air bubbles, and other irregularities in the water. While background noise in the sea limits the maximum range of detection, reverberation may limit performance in all ranges. Since reverberation originates from the signal transmission, it has a power level at any time interval closely related to that of the signal. Consequently, to detect small objects in areas where reverberation level is high (due, for example, to a shallow and rough sea bottom), the beams must be as narrow as possible in the relevant dimension. When reverberation is the limiting factor in detection, increasing the transmitted power does not improve detection over most of the range. Consequently, many sonar systems operate at acoustic powers of fairly low peak level.

Directivity. Sonar is most efficient if the sound energy is confined to a narrow beam on transmission and the receiving transducer responds only to sound coming from a limited angle, or cone. If the transmitter is highly directional, then the sound intensity at any distant point, for a given power source, will be much higher than if the sound had been broadcast uniformly in all directions. Fewer unwanted objects will be illuminated by the transmitted beam. On reception, a high directivity means that the noise will be reduced. If the noise is isotropic (coming almost equally from all directions) the reduction is given by the Directivity Index. Directivity Index is defined for transmission as the ratio in decibels of the intensity at a particular point in the direction of maximum transmission to that which would result at the same point if all the transmitted power were spread uniformly in all directions. On reception, it is the ratio in decibels of the output power developed by a signal in the direction of maximum response to that developed by the same signal power if it were uniformly distributed over all directions. For ordinary transducers, Directivity Index is the same whether the transducer is used for transmission or reception.

The Directivity Index

NEW APPLICATIONS AND TECHNIQUES

Improved sonar techniques are leading to (and often originating from) new applications in various fields of underwater operations. Since naval requirements obviously call for longer ranges of detection, higher powers and larger transducers are being provided, augmented by the sophisticated computer technique of "signal processing," whereby low-level signals can be extracted from a noise background.

Side-scan sonar. In nonmilitary applications, one technique that is increasing rapidly in importance is side-scan sonar. This sonar has a narrow (about 1°) beam in the horizontal plane, but relatively wide (10° or 20°) in the vertical plane, looking sideways from the ship or towed body on which it is mounted. As the ship proceeds along its route, a map of the acoustic scattering from the seabed is built up on recording paper. With experience, these records can be interpreted in geological terms, giving in effect a map of the hardness and roughness of the seabed rocks; it reveals geological faults, sand waves, and ridges. The general arrangement is shown in Figure 2.

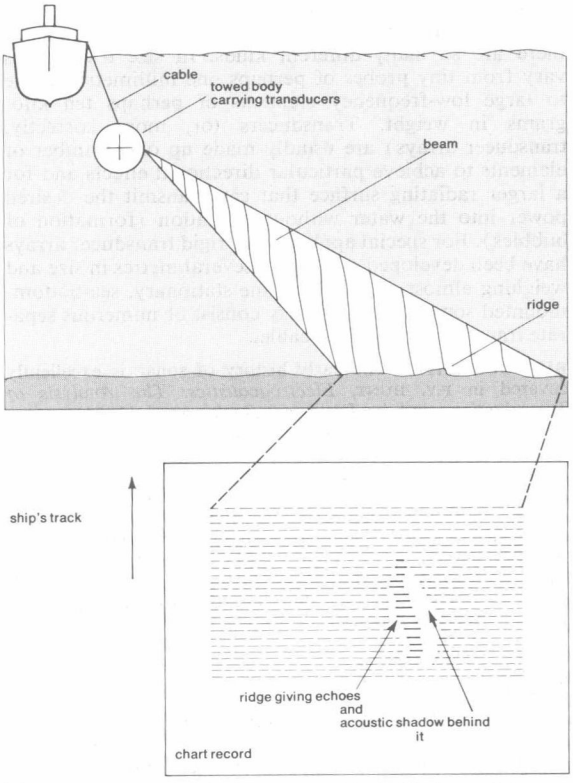


Figure 2: Principle of side-scan sonar with simplified record showing single small ridge with acoustic shadow behind.

Existing equipment for geological applications has a range of about 1,000 metres, but a recent model constructed by the British National Institute of Oceanography has a range of approximately 15 kilometres working at a relatively low frequency, of 6.6 kilohertz. The speed of search is high, but the results are seriously affected by refraction. Smaller, high-frequency, side-scan sonars, operating at 300 kilohertz (or even higher) with a range of about 200 metres, survey an area in finer detail. They can delineate such objects as underwater pipelines, oyster beds, and wrecks.

WPESS sonar. The growing trend towards higher resolution in sonars has led to the development of what is called within-pulse electronic-sector-scanning sonar (WPESS sonar). When very narrow sonar beams are used and it is desired to search a sector rapidly by swinging the beams around, mechanical beam swinging slows the rate of search, making the method unsatisfactory for fish finding and mine hunting when the ship is moving forward. In the WPESS method, a wide sector is covered by each transmitted pulse, and a narrow receiving beam is rapidly swung electronically across this sector. The beam is swung so rapidly that the whole sector width is examined for echoes before the acoustic pulse has travelled its own length through the water. In this way, the whole sector is examined on every transmitted pulse. Thus the search rate is very high, even if very narrow receiving beams are used to provide high resolution. The display is usually a cathode-ray tube giving some kind of plan view of the sector, as illustrated in Figure 3.

From *Underwater Journal* (September 1970)

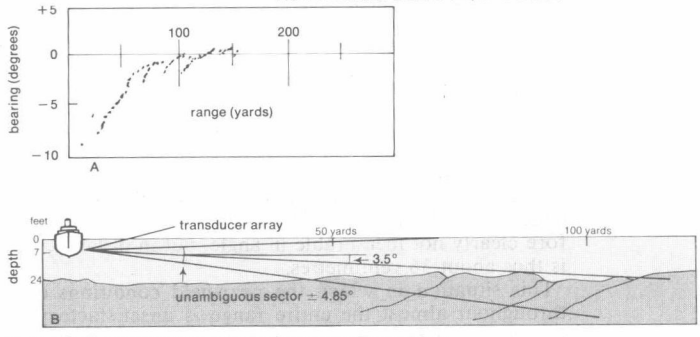


Figure 3: *Scanning sonar used with vertical 'scan' in a shallow reservoir.*
(A) Bearing-versus-range display.
(B) Geometry of bottom of reservoir deduced from A (vertical section).

Applications of WPESS sonar

The use of WPESS sonar is beneficial in the study of fish behaviour, in observing the behaviour of underwater gear, and in monitoring and directing the movements of divers. For some of these applications the acoustic frequency is made high (*e.g.*, 500 kilohertz) so that high resolution (*e.g.*, 0.5° in angle, 7.5 centimetres [three inches] in range) can be obtained using only small transducers. The maximum range is then relatively small, perhaps only 60 metres on 500 kilohertz equipment, or 200 metres in 300 kilohertz equipment. On the other hand, the system can be used equally well at lower frequencies. In early trials, 37 kilohertz was used with a range of detection on small fish shoals of nearly 800 metres.

Because of the trend to higher resolution, sonar can compete with optical methods for underwater viewing. Sonar makes it possible to view objects in muddy and turbid water where optical methods are ineffective. It is thus of great importance in police searches, in civil engineering (*e.g.*, river and harbour works), and in studying fish behaviour. In these applications it is essential to use a sonar that provides high angular (or lateral) resolution as well as high resolution in range. Range resolution can usually be increased by using a short pulse. Lateral resolution, however, raises difficulties.

Beam width considerations. Obtaining a beam of narrow angular width requires a transducer with dimensions equivalent to many wavelengths of the sonar carrier frequency. For example, a beamwidth of half a degree requires transducer dimensions of approximately 120 wavelengths. To keep the transducer small, a high frequency is necessary. Attenuation in water, however, increases rapidly with frequency. For example, at 500 kilohertz, it is in the region of 0.1 decibel per metre, while at one megahertz it is in the region of 0.3 decibels per metre. Thus the frequency cannot be increased indefinitely, and there is a limit to the reduction in size of the transducer. If, for example, the maximum range is to be about 50 metres, then 500 kilohertz would be a reasonable choice of frequency, the wavelength would be 0.3 centimetre, and the transducer length (l) for a 0.5° beam would be 36 centimetres (14 inches).

Far-field beamwidth

The concept of the 0.5° beam is based on the idea that the measuring point is far removed from the transducer from which the beam boundaries are drawn (far-field condition). But at close ranges the concept of angular beamwidth is invalid; if the transducer is straight (*i.e.*, unfocussed) the beam can hardly be narrower than the transducer length l, and the geometry becomes as shown in Figure 4. The far-field beamwidth θ is realized only after a range of about l^2/λ; *i.e.* 40 metres from the trans-

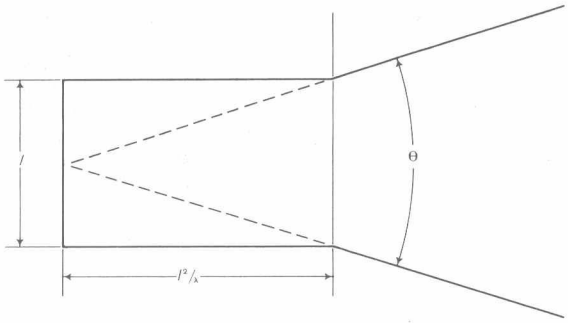


Figure 4: Geometry of beam near transducer (see text).
From B.K. Gazey and D.G. Tucker, *Applied Underwater Acoustics* (1966); Pergamon Press; reprinted with permission

ducer (λ is the wavelength in metres). This is almost the maximum range specified, so that "near-field" conditions exist almost throughout. The lateral resolution is therefore clearly not measurable in angle, but in distance, and is thus about 36 centimetres.

This situation in which the near-field conditions exist throughout almost the entire range is unsatisfactory. It can be overcome by using a system of signal processing, in which the receiving transducer is divided in half, and the signals from the two half-transducers are multiplied together. Thus a target can give a signal output from the receiver only if it lies in the beams of both half-transducers. Moreover, the far-field beamwidth is one-half of that corresponding to the length of the transducer used normally, and is realized at a relatively short distance from the transducer. Thus the near-field beam is extremely narrow, and lateral resolution of one-sixth of the transducer length (or less) is feasible. In the example taken above, the overall length of the transducer would be 18 centimetres, and the lateral resolution is quoted as three centimetres, or 0.5°, whichever is the poorer. This provides an enormous advantage for the multiplicative system.

Influence of modern electronics. Older forms of electronics with thermionic tubes or discrete transistors have been expensive to produce, somewhat unreliable, and costly to maintain. Because of the revolution in electronics that brought forth microelectronics, or integrated-circuit technology, however, it is now possible for a system to be sophisticated and complex and, at the same time, relatively inexpensive and reliable. This change in electronics has influenced sonar design, and already more refined and useful sonars are going into commercial production.

The microelectronic units that first became available were of the digital type; that is, they represented circuits that had switching, or on-off, functions of the type used in digital computers.

Several types of analogue circuits later became available in microelectronic form, and it is possible to make even a complex analogue system such as the WPESS sonar on this basis.

PRINCIPLES OF TRANSDUCERS

The purpose of transducers (strictly, electroacoustic transducers) is to convert an electrical signal into an acoustic signal, or vice versa. Three types of transducer are employed in underwater applications: magnetostrictive, piezoelectric, and electrostrictive.

Magnetostrictive. In a magnetostrictive type, a magnetic field is applied to a piece of suitable magnetic material, causing the dimension of the piece to decrease along the axis parallel to the field. When the field is alternating, application of a steady magnetic field (polarization) is necessary to give the acoustic wave the same frequency as that of the electrical signal. The magnetostrictive effect also operates in reverse: received acoustic signals cause compression of the material, altering the magnetic field, which in turn produces an electromotive force in the electrical winding.

Piezoelectric. Piezoelectric transducers use crystals that change in dimension according to the applied electric field. If the field is alternating, the crystals vibrate and radiate an acoustic wave. Conversely, if the crystals are acted on by acoustic waves, they generate an electric field. Piezoelectric materials used include quartz, ammonium dihydrogen phosphate, tourmaline, and lithium sulfate (see also PIEZOELECTRIC DEVICES).

Electrostrictive. Electrostrictive transducers are becoming the most widely used of the three. Materials used include barium titanate and lead zirconate. The change of dimensions depends on the magnitude but not the polarity of the applied electric field, so polarization is needed as with magnetostrictive transducers. Electrostrictive transducers generally have impedances of a few hundred ohms.

Transducers usually are operated at resonance; that is, the frequency at which they vibrate naturally, to get reasonable efficiency, making their frequency bandwidth rather narrow. When wide-band operation is required, the transducer is operated in such a way that its natural resonances are outside the operating frequency band. Various methods are used to reduce the sharpness of resonance as well as the cost of transducers. In one method, slices of crystal or ceramic material are alternated with plates of metal; such "sandwich" transducers are common.

It is difficult to generalize about transducer design since there are so many different kinds. In size transducers vary from tiny probes of perhaps one millimetre square to large low-frequency elements of perhaps ten kilograms in weight. Transducers (or, more correctly, transducer arrays) are usually made up of a number of elements to achieve particular directional effects and for a larger radiating surface that can transmit the desired power into the water without cavitation (formation of bubbles). For special applications, rigid transducer arrays have been developed measuring several metres in size and weighing almost a ton. For some stationary, sea-bottom-mounted sonars the array may consist of numerous separate transducers linked by cable.

BIBLIOGRAPHY. The early history of sonar is excellently covered in F.V. HUNT, *Electroacoustics: The Analysis of Transduction, and its Historical Background* (1954); sonar occurring in nature is treated in W.N. KELLOGG, *Porpoises and Sonar* (1961); and D.R. GRIFFIN, *Listening in the Dark: The Acoustic Orientation of Bats and Men* (1958). An elementary textbook is D.G. TUCKER, *Underwater Observation Using Sonar* (1966); more advanced textbooks include: D.G. TUCKER and B.K. GAZEY, *Applied Underwater Acoustics* (1966); J.W. HORTON, *Fundamentals of Sonar* (1957); and R.J. URICK, *Principles of Underwater Sound for Engineers* (1967). Sophisticated sonar techniques are described in a simple way in D.G. TUCKER, *Sonar in Fisheries: A Forward Look* (1967).

(D.G.T.)

Sonata

Deriving from the past participle of the Italian verb *sonare*, "to sound," the term sonata originally denoted a composition played on instruments, as opposed to one that was *cantata*, or "sung," by voices. Its first such use was in 1561, when it was applied to a suite of dances for lute. The term has since acquired other meanings that can easily cause confusion. It can mean a composition in two or more movements, or separate sections, played by a small group of instruments, having no more than three independent parts. Most frequently it refers to such a piece for one or two instruments. By extension, sonata can also refer to a composition for a larger instrumental group having more than two or three parts, such as a string quartet or an orchestra, provided that the composition is based on principles of musical form that from the mid-18th century were used in sonatas for small instrumental groups. The term has been more loosely applied to 20th-century works, whether or not they rely on 18th-century principles.

Sonata and sonata form

Quite distinct from all of the preceding, however, is the use of the term in "sonata form." This denotes a particular form or method of musical organization normally used within instrumental sonatas, string quartets, and other chamber music, and symphonies written since the beginning of the Classical period (the period of Mozart, Haydn, and Beethoven) in the mid-18th century.

The first concern of this article will be to establish the

principles of musical form often associated with sonatas for small and large groups of instruments. They will be approached through an examination of the principle of musical structure called "sonata form." A historical account of the origins and development both of the instrumental sonata and of sonata form will then endeavour to throw light on other meanings of the term. In conclusion, some estimate will be offered of the present and possible future roles of the sonata in musical life.

STRUCTURE OF THE CLASSICAL SONATA

Sonata form denotes a particularly fertile manner of organizing the musical structure of a single movement. It commonly occurs within the larger context of a multimovement scheme. Maturing in the second half of the 18th century, it provided the instrumental vehicle for much of the most profound musical thought until about the middle of the 19th century, and has continued to figure largely in the methods of composers down to the present day.

The basic elements of sonata form are three: exposition, development, and recapitulation, in which the musical subject matter is stated, explored or expanded, and restated. There may also be an introduction, usually in slow tempo, and a coda, or tailpiece, but these optional sections do not affect the basic structure. Although sonata form is sometimes called first-movement form, the first movements of multimovement works are not always in sonata form, nor does the form occur only in first movements. Likewise, another name for it, sonata-allegro form, is misleading, for it need not be in a quick tempo such as allegro.

At first glance sonata form may appear to be a species of three-part, or ternary, form. The three parts of ternary form are a first section (A), followed by a contrasting section (B), followed by a repetition of the first section (that is, ABA). The parts are interrelated not in terms of basic structure but by purely lyrical or character contrast. Actually, the three parts of sonata form developed out of the binary, or two-part, form prominent in the music of the 17th and early 18th centuries. In binary form the structure depends on the interrelationship not only of themes but also of tonalities, or keys, the particular sets of notes and chords used in each part. Thus, the initial part, which is repeated, leads directly into the second part by ending in the new key in which the second part begins. The second, also repeated, moves from the new key back to the original key, in which it ends. The second part thus completes the first.

Binary form

In sonata form the exposition corresponds to the first part of binary form, the development and recapitulation to the second. The exposition moves from the original key to a new key; the development passes through several keys and the recapitulation returns to the original key. This echoes the motion, in binary form, away from and back to the original key. In relation to binary form, sonata form is complex. It offers, in the exposition, contrasting musical statements. In the development these are treated dialectically; that is, they are combined, broken up, recombined, and otherwise brought into change and conflict. In the recapitulation they are restated in a new light. This organic relationship between parts marks the sonata form as a higher, more complex, type than the ternary form. The occasional designation of sonata form as compound binary form is useful in that it stresses its origins in the earlier form, but notes its added complexity.

The exposition. The emphasis on contrast, even conflict, is the element that distinguishes the exposition of a sonata-form movement from the first section of an earlier binary form. The first section of a binary movement in a Baroque suite or instrumental sonata, for example, might contain two clearly differentiated themes, but the stress is on continuity and on uniformity of musical texture rather than on contrast. In sonata form the emphasis is more dynamic; there is a stronger sense of contrast within the movement. The terms usually given the contrasting areas are "first subject/second subject" or "principal group/subsidiary group." These are misleading terms, for they imply a simple contrast of themes.

In reality it is contrast of key, or tonal contrast, that characterizes the sonata exposition. Usually the opening of the exposition is firmly rooted in the tonic, or "home," key of the work. The later segments of the exposition move decisively to a closely related but distinct key. The second key chosen is almost invariably one of the two keys most closely related to the home key. If the home key was a major key, the dominant key is chosen; if the home key was minor, the relative major is chosen. (The dominant key is the one whose keynote is five tones above that of the tonic, as C–G; the relative major has a keynote three tones above the relative minor, as A minor–C major.) The exposition thus creates an opposition of tonalities or key areas that the rest of the movement—the development and recapitulation—will strive to reconcile. Compared with the contrast of keys, the question of how many themes the movement possesses is of minor structural significance. Very often, a movement in sonata form has two clearly defined main themes, for example the first movement of Mozart's *Symphony No. 41 in C Major*. It may also have only one, like the first movement of Haydn's *Symphony No. 85 in B Flat Major*. Or it may have more than a half dozen strongly characterized themes, as does the first movement of Brahms's *Fourth Symphony*.

Thematic and tonal contrast

The thematic organization of a movement in sonata form may affect the character of the exposition, and thus of the whole movement, in two specific respects. When two themes or groups of themes are clearly differentiated, their distribution may help the listener to assimilate the cardinal points of the tonal design (that is, the arrangement of keys) of the movement. When, on the other hand, such differentiation between themes is obscured or set at variance with the organization of tonalities, the very tension between thematic design and tonal scheme may greatly enhance the subtlety and interest of the form. Such tension may produce not merely an interplay of melody and key within the movement but an interplay between two interplays. One fairly simple way of achieving this is shown in the first movement of Haydn's *Symphony No. 99 in E Flat Major*. Here, as in *No. 85*, the first theme is restated in the dominant key. This restatement could appear at first to be the second subject. But later it is followed by another distinct motif that, in terms of themes, is the real second subject. At the same time the neat, almost epigrammatic character of the second subject makes it similar to a codetta theme, which is often used to round off the exposition after both main subjects have been stated.

In the first movement of Mozart's *Symphony No. 35 in D Major (Haffner)* this interplay of interplays reaches a higher level of subtlety. The second theme, against which the first persists as a counterpoint, is stated "on" rather than "in" the dominant; that is, its harmonies suggest the dominant key, but remain part of the home, or tonic, key. The second theme is thus heard as a new perspective on the tonic. Later, when the dominant key is firmly established in its own right, Mozart introduces a new subject whose tune is closely related to the first theme. In this richly ambiguous structure, the newly introduced motif would be regarded by the criterion of key, as the second subject; in purely thematic terms, it might almost be said to constitute the beginning of the codetta, or concluding section.

The development and the recapitulation. The functions of the other two main sections follow naturally from what has been established in the exposition. Their purpose is to discuss and resolve the conflicts of tonality and theme that the exposition has raised. The development is an area of tonal flux—it usually modulates, or changes key, frequently, and any keys it settles in are likely to be only distantly related to the keys found in the exposition. It frequently proceeds by breaking the principal themes down into smaller elements and bringing these elements into new tonal or contrapuntal relations with each other. That is, themes or fragments of themes may appear in new keys; they may be combined to form apparently new melodies; they may be played against each other as counterpoint, or countermelody. One of the finest illustrations

Modulation and thematic alteration

of the methods of development used in the Classical period occurs in the first movement of Mozart's *Symphony No. 38 in D Major* (*Prague*). Another resource of development is to seize on an apparently minor feature of the exposition and, by developing it extensively, to demonstrate its hidden importance. Another is to introduce entirely new material. This may provide a moment of relief in the course of a rigorous argument (as in the last movement of Mozart's *Piano Sonata in C Major*, K. 330); or it may allow the composer to expand the scope of a large-scale movement (as in the first movement of Beethoven's *Symphony No. 3 in E Flat Major*, the *Eroica*).

Sometimes such a theme may only *seem* to be new. In the first movement of Beethoven's *Symphony No. 4 in B Flat Major*, for instance, the theme in the development that is usually described as "new" is really a decorated version of a motif already heard in the exposition.

One common tactic in the Classical development is to begin the section with the codetta theme that ended the exposition. The first movement of Beethoven's *Cello Sonata No. 2 in G Minor*, Opus 5, No. 2, is an example.

The impact of this device, and of the development section as a whole, is often obscured by the common tendency among modern performers to ignore the composer's instruction, present in almost all sonata form movements of the Classical period, to repeat the entire exposition. When this repetition is omitted, the thematic balance of the movement is upset and the dramatic effect of the development's sudden departure from an established regularity can be ruined. Music is an art to which the controlled use of time is basic. The temporal structure of a movement cannot be altered without seriously changing the proportions of the whole.

Tension and recapitulation

Like the beginning of the development section, the point at which development passes into recapitulation is one of the most important psychological moments in the entire sonata structure. It marks the end of the main argument and the beginning of the final synthesis for which that argument has prepared the listener's mind. The Classical masters differ interestingly in their handling of this juncture. All of them usually prepare for it with a long passage of gathering tension. In Mozart the return of the tonic key and subject is managed with understated punctuality, the actual moment of recapitulation gliding in almost unnoticed. Haydn and Beethoven tend to celebrate its advent with panoply.

The recapitulation presents the principal subject matter of the movement in a new state of equilibrium. The main subjects of the exposition are heard almost always in the same order as before, but now both subjects are typically in the tonic key, whereas in the exposition the first was in the tonic, the second in the dominant key. As a result of the musical events in the development, the listener perceives the subjects in a new relationship—rather like a traveller who glimpses the constituent parts of a valley separately as he climbs a hill and then, when he reaches the summit, sees the entire landscape for the first time as a whole. The recapitulation can vary greatly in the literalness with which it repeats the elements of the exposition. Sometimes, as in the first movement of Mozart's *Sonata in B Flat Major*, K. 570, a tiny modification in the transition that originally led from the tonic to the dominant key is enough to effect the necessary change of key perspective and keep the second subject in the tonic key. In other cases (the first movement of Beethoven's *Eroica Symphony*, for instance, and many of Haydn's symphonic movements) far-reaching modifications and reshufflings of the original material are made in the recapitulation. As in any living manifestation of a principle of musical form, the methods differ vastly from work to work; but the effect is always to bring about the reconciliation of opposites that is essential to sonata form.

A large-scale sonata movement often creates conflicts of key and theme that cannot be completely settled even by the full process of recapitulation. In this case, the movement may be rounded off with a coda, or concluding section. Beethoven often extends the coda so greatly that it becomes almost a second development section, as in his *Appassionata* piano sonata. But this is no more an essential element of sonata form than the introduction that may precede the main movement.

Interrelationship of movements. The form described above is that exemplified in most first movements of sonatas, and of sonata-style compositions, in the Classical period. There are usually two, three, or four movements in the entire work. Two-movement and, more particularly, three-movement schemes are most common in sonatas for one or two instruments. Symphonies and string quartets almost always have four movements, and Beethoven, particularly in his earlier period, sometimes expanded the scheme of the instrumental sonata to four movements too.

The first movement in all of these patterns is usually fast; the second commonly provides the contrast of a slower tempo; and the last in most cases is again fast. When there are four movements, a simpler, dance-style movement of the type also found in the suite is included. This is usually placed between the slow second movement and the finale; in some cases it stands second and the slow movement third.

The forms of these other movements vary much more than that of the first, which in Classical examples is almost invariably the weightiest. Since their function is to complement the experience of the first movement through a new but related range of contrasts, their scope and manner depend on the point to which the issues of the work have already been taken. Simple ternary (ABA) form and variation form (*i.e.*, theme and variations) are among the most common patterns for the slow movement, but rondo and sonata forms are also used. In rondo form a recurring theme is contrasted with a number of intervening themes, as ABACA. When sonata form is used in slow tempos, the demands of overall proportion frequently cause the omission of the development section. Sonata form, rondo, and, less often, variation form are also used for the final movement. In final movements, also, the simple rondo pattern (ABACA) is often expanded into ABA-development-BA, with B in the dominant key at its first appearance and in the tonic key at its second. The result is a hybrid form known as sonata-rondo.

The dance movement

In the first part of the Classical period, the dance movement, when it appeared, usually consisted of a minuet in fairly simple binary form (the two-part form from which sonata form evolved). This was followed by a second minuet known as the trio, which tended in orchestral works to be more lightly scored. The first minuet was then repeated, normally without its own internal repeats. The minuet-trio-minuet structure forms an overall ternary pattern. Haydn frequently, and Beethoven still more often, chose to speed the traditional minuet up to the point at which it lost its dance character and became a scherzo, a quick, light movement usually related to the minuet in form. In some extreme cases, such as the ninth symphonies of both Beethoven and Schubert, the binary structures of both scherzo and trio were expanded into small but complete sonata-form structures. In this way, as with the sonata-rondo, the principles of thematic development and key contrast spread during the Classical period as the sonata form began to influence other movements.

Such are the outlines of the most fertile form in Western instrumental music since 1750. Before discussing its origins, growth, and later modification, there should be a warning against phrases like "true sonata form." If sonata form as described is considered the sole "true sonata form," the implication is that instrumental sonatas written in forms other than sonata form are not genuine.

Actually, the principle exemplified in sonata form represents one way of organizing the passage of sound through time. Its scope is enormous; it was the basis for some of the greatest works of Western music; and it still contains the seeds of potential further development. But it is only one episode in a complex chronicle of styles and principles of musical organization. In contrast to earlier forms, it emphasizes conflict instead of continuity and derives its impact from the explosive power of tonal organization instead of the smoother influence of melody.

THE SONATA IN MUSICAL HISTORY

The sonata in all its manifestations has roots that go back long before the first uses of the actual name. Its ultimate

Earliest origins

sources are in the choral polyphony of the late Renaissance (music having several equal melodic lines, or voices). This in turn drew at times on both liturgical and secular sources—on the ancient system of tones or modes of Gregorian chant, and on medieval European folk music. These two lines were constantly interweaving. Popular tunes, for example, were used as the starting point for masses and other religious compositions from the 15th to the early 17th centuries. Sacred and secular elements influenced the development of both the sonata and the partita (or suite) of the Baroque period.

The specific musical procedures that were eventually to be characteristic of the sonata began to emerge clearly in works by the Venetian composers of the late 16th century, notably Andrea Gabrieli (*c.* 1520–86) and Giovanni Gabrieli (1556–1612). These composers built instrumental pieces in short sections of contrasted tempo, a scheme that represents in embryo the division into movements of the later sonata. This approach is found not only in works entitled "sonata," such as Giovanni Gabrieli's *Sonata pian' e forte* (*Soft and Loud Sonata*) of 1597, which was one of the first works to specify instrumentation in detail; the instrumental fantasia and the canzone, an instrumental form derived from the chanson or secular French part-song, display a similar sectional structure. Like early sonatas, they were often contrapuntal (*i.e.*, built by counterpoint, or the interweaving of melodic lines in the different voices, or parts). At this stage sonatas, fantasias, and canzoni were often indistinguishable from each other, and from the fuguelike ricercare, though this form is generally more serious in character and more strictly contrapuntal in technique.

In the 17th century stringed instruments eclipsed the winds, which had played an at least equally important role in the sonatas and canzoni composed by the Gabrielis for the spacious galleries of St. Mark's Cathedral, Venice. Claudio Monteverdi (1567–1643) devoted more of his energies to vocal than to instrumental composition. The development of instrumental writing—and of instrumental musical forms—was carried on more and more by virtuoso violinists. One of these was Carlo Farina (fl. *c.* 1630), who spent part of his life in the service of the court of Dresden, and there published a set of sonatas in 1626. But the crowning figure in this early school of violinist-composers was Arcangelo Corelli (1653–1713), whose published sonatas, beginning in 1681, sum up Italian work in the field to this date.

Apart from their influence on the development of violin technique, reflected in the works of such later violinist-composers as Giuseppe Torelli (1658–1709), Antonio Vivaldi (1678–1741), Francesco Maria Veracini (1690–*c.* 1750), Giuseppe Tartini (1692–1770), and Pietro Locatelli (1695–1764), Corelli's sonatas are important for the way they clarify and help to define the two directions the sonata was to take. At this point the *sonata da chiesa*, or church sonata, and the *sonata da camera*, or chamber sonata, emerged as complementary but distinct lines of development.

Church and chamber sonatas

The *sonata da chiesa* usually consists of four movements, in the order slow–fast–slow–fast. The first fast movement tends to be loosely fugal in style (*i.e.*, using contrapuntal melodic imitation), and thus reflects, most clearly of the four, the sonata's roots in the fantasia and canzone. The last movement, by contrast, is simpler and lighter, often differing from the dance style typical of the *sonata da camera* only in that its sections are not repeated. The *sonata da camera* is altogether less serious and less contrapuntal than the *sonata da chiesa*, and it tends to consist of a larger number of shorter movements in dance style. If the *sonata da chiesa* was the source from which the Classical sonata was to develop, its courtly cousin was the direct ancestor of the suite, or partita, a succession of short dance pieces; and in the 18th century, the terms suite and partita were practically synonymous with *sonata da camera*. The two streams represented by church and chamber sonatas are the manifestation, in early Baroque terms, of the liturgical and secular sources found in Renaissance music. The Baroque style flourished in music from about 1600 to about 1750. Down to the middle of the 18th century the two influences maintained a high degree of independence; yet the injection of dance movements into the lighter examples of the *sonata da chiesa* and the penetration of counterpoint into the more serious suites and *sonate da camera* show that there was always some cross-fertilization.

Another characteristic of the Baroque sonata that Corelli's work helped to stabilize was its instrumentation. Around 1600 the musical revolution that began in Italy had shifted emphasis from the equal-voiced polyphony of the Renaissance and placed it instead on the concept of monody, or solo lines with subordinate accompaniments. The comparatively static influence of the old church modes was superseded by the more dramatic organizing principle of the major–minor key system with its use of contrast of keys. Although counterpoint continued to play a central role in musical structure for another hundred years and more, it became a counterpoint that took careful account of the implications of harmony and of chords within the framework of the major and minor keys.

The continuo part

In this context the continuo, or thorough bass, assumed primary importance. The composer that used a continuo part wrote out in full only the parts of the upper melody instruments. The accompaniment, which was the continuo part, was given in the form of a bass line, sometimes supplemented with numbers, or figures, to indicate main details of harmony, whence the term "figured bass." The continuo was "realized," or given its performed form, by a low melody instrument (viola da gamba, violone, or later cello or bassoon) in collaboration with an organ, harpsichord, or lute. The collaborating instrument improvised the harmonies indicated by the figures or implied by the other parts and so filled the gap between the treble and bass lines.

In Corelli's work, "solo" sonatas, for one violin with continuo, are found alongside others for two violins and continuo described as sonatas *a tre* (*i.e.*, for three), early examples of the trio sonata that was the principal chamber-music form until about 1750. The use of "trio" for sonatas played by four instruments is only superficially paradoxical: although trio sonatas were played by four instruments, they were considered to be in three parts—two violins and continuo. Moreover, specific instrumentation at this period was largely a matter of choice and circumstance. Flutes or oboes might play the violin parts, and if either harpsichord or cello or their substitutes were unavailable, the piece could be played with only one of them representing the continuo. But a complete continuo was preferred.

Corelli's importance is as much historical as musical. Perhaps because a vigorous line of Italian composers of violin music followed him, he is commonly accorded the main credit for late 17th-century developments in sonata style. But his undeniably vital contribution should not distract attention from equally important work that was done around the same time outside Italy.

Early development outside Italy

In France Jean-Baptiste Lully's lucrative monopoly of music at the royal court and the immense popularity of spectacular ballets used as courtly entertainments naturally led, through François Couperin (1668–1733), to a concentration on the smaller dance forms found in the ballet and courtly social dance. This concentration gave the French school its pre-eminence as producer and influencer of the 18th-century dance suite. The French, thus occupied with dance music, had little effect on the growth of the *sonata da chiesa*. But in Germany, where in 1619 Michael Praetorius (1571–1621) published some of the earliest sonatas, the sonata developed from an originally close relation to the suite into a more ambitious blend. As it evolved it combined the suitelike multisectional structure of the *sonata da camera* with the contrapuntal workmanship and emotional intensity of the Italian *sonata da chiesa* form.

One of the first contributors to this development of the Italian influence was the Austrian composer Johann Heinrich Schmelzer (*c.* 1623–80). In Nürnberg in 1659 he published a set of trio sonatas for strings, following it in 1662 with a set for mixed strings and wind instruments, and in 1664 with what may have been the first set of so-

natas for unaccompanied violin. The German composer Johann Rosenmüller (*c.* 1620–84) spent several years in Italy; his *Sonate da camera cioè sinfonie* (*i.e.*, suites or symphonies), published in Venice in 1667, are essentially dance compositions. But 12 years later, in Nürnberg, he issued a set of sonatas in two, three, four, and five parts that vividly illustrate the German trend toward more abstract musical structure and expressive counterpoint. During this period even pieces with dance titles began to lose their danceable character and became compositions meant only for listening.

Meanwhile, the greatest member of this school, Heinrich Ignaz Franz von Biber (1644–1704), published several sets of sonatas—some for violin and continuo, others in three, four, and five parts. In these, from 1676 onward, he took a penchant for expressiveness to extremes of sometimes bizarre but often gripping profundity that contrast sharply with the bland, polished style of Corelli. The titles of some of Biber's sets of sonatas specifically indicate his aim of reconciling church and chamber styles. The 1676 publication, for instance, is entitled *Sonatae tam aris quam aulis servientes* (*Sonatas for the Altar as Well as the Hall*). And being himself, like Corelli, a violinist of extraordinary powers, Biber made a valuable contribution to the development of instrumental technique in a set of sonatas for unaccompanied violin in which the practice of *scordatura* (adjustment of tuning to secure special effects) is ingeniously exploited.

The English composers were achieving a comparable intensification of expression during the 17th century, though in their case the technical starting point was different. In accordance with the characteristic time-lag of the English in the adoption of new European musical methods, the English continued to work with polyphony in the Renaissance manner, while the Italians were perfecting monody and the Germans fruitfully uniting monody with their own contrapuntal tradition. English polyphony in the 17th century attained a remarkable level of technical finish and emotional grandeur. Thomas Tomkins (1572–1656), Orlando Gibbons (1583–1625), John Jenkins (1592–1678), and William Lawes (1602–45) were the chief agents of this refining process. They and their predecessors, notably John Coperario (*c.* 1575–1626), made a gradual transition from the string fantasia bequeathed by William Byrd and other composers during the reign of Elizabeth I (1558–1603) and approached the new kind of musical form associated with the Baroque sonata; but they always stayed closer than their continental colleagues to the spirit of polyphony.

Contribution of English polyphony

When Henry Purcell (*c.* 1659–95), in his three-part and four-part sonatas, submitted this rich English tradition to the belated impact of French and Italian influence, he produced a fusion of styles that was the highest point of musical inspiration yet reached by the emergent sonata form.

The Baroque sonata. The years from the end of the 17th century to the middle of the 18th represent a moment of equilibrium in the interaction of counterpoint and monody that had created the Baroque sonata. The continuo device, as long as it endured, was a sign that the balance still held—and it did endure as long as the trio sonata kept its central position as a chamber-music medium. During the first half of the 18th century the later Italian violinists, most notably Vivaldi, were prolific creators of trio sonatas. Sometimes they leaned to a three-movement pattern (fast–slow–fast), influenced by the direction the Italian operatic sinfonia, or overture, was taking. More often the old four-movement pattern was preserved. In this well-tested shape, too, Georg Philipp Telemann (1681–1767) produced hundreds of examples that maintained a remarkably consistent standard of musical interest, George Frideric Handel (1685–1759), working for most of his life in England, composed some trio sonatas, and also some valuable sonatas for solo instrument with continuo. In France, Joseph Bodin de Boismortier (1691–1755) and the violinist Jean-Marie Leclair the elder (1697–1764) cultivated both solo and trio genres with charm although with less profundity.

Yet even while the sonata with continuo flourished, the forces of tonality, or organization in terms of keys, developed intensely toward a use of key contrast that would eventually drive the trio sonata from the scene. The continuo itself was being undermined by the growth of interest in instrumental colour, and the figured bass could not long survive the tendency toward scoring for specific instruments and exhaustive detailed musical notation.

Rise of tonal consciousness

Beginning in 1695 Johann Kuhnau (1660–1722) had published the first sonatas for keyboard instrument alone, some of them programmatic pieces on biblical subjects. J.S. Bach (1685–1750), the greatest composer of Baroque sonatas, continued the move away from the treatment of the keyboard in the subordinate, "filling-in" capacity that was its role in the continuo. He wrote a small number of trio sonatas after the traditional scheme, and also a few violin and flute sonatas with continuo; but at the same time he produced the first violin sonatas with obbligato harpsichord parts (that is, obligatory and fully written out, rather than improvised), others for flute or viola da gamba with obbligato harpsichord, and three sonatas (along with three partitas) for unaccompanied violin.

In these works, as in some of Telemann's later sonatas, the power of key or tonality to articulate sections of musical structure, and its ability to provide a harmonically derived eventfulness—a sense of expectation succeeded by fulfillment—began to make itself felt. These powers of key are the seed from which the Classical sonata form originated. But at this point the dualism engendered by tonal and thematic contrast had not yet supplanted the more continuous, unitary processes at work in a composition based on counterpoint. Nor was the consciousness of tonality any more advanced in the otherwise forward-looking work of Domenico Scarlatti (1685–1757). His harpsichord sonatas—555 movements survive, many designed to be played in pairs or in groups of three—are often original to the point of idiosyncrasy in expression. They introduced a valuable new flexibility in the treatment of binary form, and they had a powerful effect on the development of keyboard writing. But in formal terms they still belong in the old world of unity—even their strongest contrasts have an air of being suspended in time, quite unlike the far-ranging effects of conflict through time that are the basis of the Classical sonata.

A later generation of composers completed the transition from Baroque to Classical sonata. One of J.S. Bach's own sons, Carl Philipp Emmanuel Bach (1714–88), plunged enthusiastically into the new resource of dramatic contrast. In about 70 harpsichord sonatas, and in other works for chamber ensembles and for orchestra, he placed a new stress on key contrast not only between but, more important, within movements. Correspondingly, he emphasized the art of transition.

Transition from Baroque to Classical

In the development of sonata form in orchestral music, particular value attaches to the work of the Austrians Georg Matthias Monn (1717–50) and Georg Christoph Wagenseil (1715–77) and of the Italian Giovanni Battista Sammartini (1701–75). All three played vital roles in shaping the symphony, which assumed an importance equal to that of the solo or small-ensemble sonata. Their symphonies further stressed the individual characterization of themes and, in particular, the use of the second subject to shape form. Another of Bach's sons, Wilhelm Friedemann Bach (1710–84), made sporadic but interesting contributions to this development, and a third, Johann Christian Bach (1732–82), who settled in London, exploited a vein of melodic charm that influenced Mozart.

The Classical sonata and after. By about 1770 most of the specific changes that dictated the shift from Baroque sonata to Classical sonata were firmly established. Through the work of the Neapolitan school of opera led by Domenico Scarlatti's father Alessandro (1660–1725), the operatic sinfonia, or overture, had streamlined the traditional *sonata da chiesa*. It omitted the opening slow movement and abandoned the fugal manner that was the first allegro's link with the past. In the new three-movement pattern, a minuet sometimes replaced the fast, abstract finale. In other cases, the inclusion of both minuet and finale brought the number of movements back to four. The south German Mannheim school of composers

—most notably Johann Wenzel Stamitz (1717–57) and his son Karl (1745–1801)—developed the technique of the orchestra, whose resources now provided an ideal laboratory for experimentation with the dramatic effects of tonal contrast.

By this time the Classical sonata proper (*i.e.*, with at least one movement in sonata form), whether in the medium of sonata, trio, quartet, quintet, or symphony, could provide a vehicle for consolidating the process begun nearly two centuries earlier by the revolution from equal-voiced polyphony to monody, with its emphasis on melody and harmony. The Rococo style of the mid-18th century, generally known as style galant, had attained a halfway stage in which counterpoint had been virtually dropped and tunes had occupied the forefront of interest. But now, in the mature Classical style of Haydn and Mozart, superficial melodic interest was in turn subordinated. In this style the value of tunes lay in their role as functions of tonality. Key by this time had assumed a central role as the fundamental articulator of form. As a corollary, musical themes were often, though not always, reduced to the status of mere motifs, or tags. The theme's harmonic implications, which contribute to the feeling of key, took precedence over its attractiveness as melody.

Subordination of melody to tonality

The new musical principle—that of contrast of key—reached full expression in Haydn and Mozart through their use of sonata form as a principle of musical organization. Haydn's most valuable work in the sonata form is found in his series of over 80 string quartets, over 100 symphonies, 52 keyboard sonatas, and 31 trios for piano, violin, and cello. Unlike Haydn, Mozart was at his greatest in the fields of opera and of the solo concerto. (The latter, though it shared with sonata form such elements as the central principle of key contrast, was a medium that evolved, through the Baroque concerto grosso, from the fundamentally different source of the solo vocal aria and the vocal-instrumental concerto; see CONCERTO.) But in the last six symphonies, the last ten string quartets, about a dozen keyboard sonatas, and several trios, quartets, quintets, and serenades, Mozart achieved outstanding examples of sonata structures. The formerly prominent sonata for violin plays a relatively minor part in both men's output: the violin had been eclipsed by the rise of interest in keyboard instruments. It was reintroduced almost surreptitiously as a distinctly subordinate partner and regained a leading role only toward the end of the 18th century in Mozart's later violin sonatas and then in Beethoven's.

The strikingly individual details in Haydn's and Mozart's handling of sonata form are all features consonant with the general outlines of the form. Examples of different approaches include Haydn's taste for combining dualistic key schemes with monistic thematic material (that is, the use of the same basic theme in both keys). He also frequently set slow movements in keys only distantly related to the key of the first movement. Mozart preferred strongly differentiated themes, and he often reshaped his second subjects drastically when they reappeared in the recapitulation. Beethoven, in his sonata compositions (preeminently, the 32 piano sonatas, the 16 string quartets, the trios, the 9 symphonies, and the sonatas for violin and for cello), retained the basic sonata form. But he vastly extended its scale; *e.g.*, by increasing the importance of the coda, or concluding section, and by using unusual keys in the exposition, which was greatly increased in length. He also introduced extramusical implications of a profound philosophical nature. In his later sonatas and quartets he began to move away from the dualistic sonata principle and back to the monistic approach exemplified in variation form and fugue.

The case of Franz Schubert (1797–1828) is quite different. The first movement of the *Symphony No. 5 in B Flat Major* (written when he was 19) is one among several places that illustrates a changing attitude to sonata principles. In the recapitulation of this movement, the first theme is given in the subdominant key (the key whose keynote is five tones below that of the tonic, or home key, as F–C, just as the tonic's keynote is five tones below that of the dominant key, as C–G). This device enables Schubert to place the second theme in the tonic key (the goal of the recapitulation) without altering the transition between the two themes; for the same passage that, in the exposition, took the music from the tonic to the dominant serves, in the recapitulation, to take it from the subdominant to the tonic. This essentially labour-saving procedure is evidence of a certain lack of patience with the workings of sonata form as hitherto practiced. Up to this time, sonata form, first treated as a textbook study after Schubert's death, was not a set of rules codified by theorists and followed by composers. Rather, it was a principle of composition that grew out of earlier forms and that can be generalized from an examination of the actual work of Haydn, Mozart, Beethoven, and their contemporaries.

19th-century trends

Schubert's interests lay in new directions, and the first steps in two such directions are to be found in the greatest of his instrumental works. His later sonata-form compositions in all media—when they follow the rough traditional scheme of exposition, development, and recapitulation—modify it substantially. He frequently expanded the number of tonal centres (central keys) in the exposition, and sometimes also the number of basic themes, from two to three. This tendency to expansion affects the whole subsequent course of the Austro-German symphonic tradition. It is the direct ancestor of the expositions of Anton Bruckner (1824–96), with their three distinct thematic groups, and of the vastly extended sonata structures of Gustav Mahler (1860–1911). At the same time, Schubert's *Fantasy in C Major* (*Wanderer*) for piano (1822) exemplifies an opposite 19th-century trend toward contraction, through the fusion of the sonata's formerly separate movements in one closely integrated whole: the four movements of the fantasy are based on transformed versions of a single theme. Similarly, in France, Hector Berlioz (1803–69) in his *Symphonie fantastique* transformed the theme representing the artist's beloved (the *idée fixe*, or fixed idea) so that it took different forms in each movement. In this case the transformation was affected by the program or "plot" of the symphony. This was a departure from the abstract, or plotless, character of the Classical sonata. The tendency to fusion—that is, to thematic unity between movements—was the source of the thematic transformations used in symphonic poems, such as those of Franz Liszt (1811–86), as a basic principle of musical structure. But in these works the program rather than any abstract musical form suggests the particular course of the transformation of the themes. For this reason their specific form does not depend, as did that of the Classical sonata, on the exposition-development-recapitulation principle of contrast, conflict, and reconciliation of keys. A corresponding evolution away from the Classical form of the sonata for one instrument occurs in Liszt's one-movement *Piano Sonata in B Minor* (1853). In this work he used a single extended movement with subdivisions analogous to the sections of sonata form. But the specific use of his four themes, which are transformed and combined in free fashion, departs from the usual order of the classical Sonata.

The influence of program music

Robert Schumann (1810–56) likewise experimented fruitfully, especially in his *Symphony No. 4 in D Minor*, with the Schubertian idea of fusing movements together. The tendency to use thematic transformation in a manner that moved away from the Classical sonata form was complemented by César Franck, who adhered to the basic form but from 1841 utilized a "cyclic" approach; that is, one of fusion, or thematic relationships between movements. Brahms, on the other hand, carried the more familiar Classical sonata form to its highest point of complexity. In addition to making valuable innovations in rhythmic structure, he gave the role of counterpoint a new lease of vigour and interest and used the concept of thematic relationships between movements in a particularly subtle way. Chopin's three piano sonatas are concerned more with lyrical expression than with innovative formal methods. Similarly, the sonata compositions of Carl Maria von Weber (1786–1826) and Felix Mendelssohn (1809–47), which generally followed the patterns of their Classical predecessors and were highly regarded

in their day, contributed little to the evolution of sonata form away from its Classical state. This evolution, illustrated by the works of Berlioz and Liszt, was carried forward by Mahler. In his symphonies expansion, through inclusion of more than two tonal centres and groups of themes, and fusion, or the creation of unity between movements, are combined. This gives rise to expansive compositions held together by complex interrelationships between themes. Arnold Schoenberg (1874–1951), in such works as his *First String Quartet* (1904), carried the idea of fusion of movements to its logical conclusion: this is a one-movement work with contrasting sections in which all the themes used are derived from a few basic motifs.

Modern directions. Two important 19th-century developments tended to weaken the effectiveness of the Classical sonata form as an organizing principle. One, exemplified by Richard Wagner (1813–83), was an increasing use of chromaticism; that is, of notes and chords foreign to the key in which a passage of music is written. Chromaticism, when used extensively, broke down key feeling. Instead of being heard as a contrast to, or special modification of, the key, it became so prominent that the key itself was not heard strongly enough to establish itself in the listener's mind. Secondly, Liszt and his followers weakened the sonata form by using in their symphonic poems musical organizations based on program rather than on contrast of keys. But although the effectiveness of key as a basis for musical organization was weakened by the late 19th century, Mahler and Carl Nielsen (1865–1931) provided a modification of the sonata form that made use of tonality in a new way. This innovation, progressive tonality, used the home key as a goal to be worked toward from more or less distant key regions, so that a work ends in a different key from the one in which it began. Mahler and Nielsen arrived at the same notion independently at the same time. Nielsen's *First Symphony* and Mahler's *Second Symphony* (both 1894) are the first to use progressive tonality, and both composers forged highly individual new forms from it.

Weakening of the Classical sonata form

Most compositions written in sonata form after Wagner's era, however, lack a certain sense of vitality. Frequently, because the effectiveness of key or tonality has been weakened, such compositions centre on melody without the strong contrast of tonality that underlay the Classical sonata. Some composers made stylistic compromises. Schoenberg's *Fourth String Quartet* used his twelve-tone (dodecaphonic) approach to composition, an approach that began with a "row," or series, of the 12 tones of the chromatic scale, chosen by the composer to serve as the melodic and harmonic basis for the composition. In this work he fits the twelve-tone style into the outlines of sonata form. The result is based on contrasting themes, rather than on the Classical sonata principle of key contrast, because twelve-tone music, being atonal, deliberately avoids the creation of a sense of key. In a comparable way, though in the context of a different style, some of the sonatas of Sergey Prokofiev (1891–1953) use the outward formal divisions of the classical sonata form but stress the interest of melody as such, leaving tonality —still present in this case—to play a decorative rather than a structural role.

New principles

Other modern composers developed new principles of musical form. Although these principles appear in genres traditionally associated with the sonata, such as instrumental sonatas, string quartets, and orchestral works, they vary in the degree to which they are or are not related to the Classical sonata form.

One of the more useful of such principles has been the technique of constructing large-scale compositions from transformations and developments of a single germinal motif, often merely two or three notes. Like Schoenberg's approach, in which a twelve-tone row is transformed, this is actually the application at a more radical and consistent level of the 19th-century principle of thematic transformation. The symphonies of Jean Sibelius (1865–1957) are based on this method. So are those of Ralph Vaughan Williams (1872–1958), who also used some of the features of sonata form but imaginatively reshaped them and transformed their proportions to suit his purpose. In the nonsonata works of Schoenberg and his pupils Alban Berg (1885–1935) and Anton Webern (1883–1945), the twelve-tone method produced legitimate new forms of the highest historical importance; but when forced into an uncomfortable liaison with earlier schemes of organization such as the sonata, its effectiveness diminished. In the works of Béla Bartók (1881–1945), passages built on folk music scales, rather than on the major and minor scales of 18th- and 19th-century keys or tonalities, are used alongside atonal passages. His musical structure frequently takes the form of a combination of elements of sonata form with a simple "archlike" structure such as ABCBA. Paul Hindemith (1895–1963) contributed copiously to the sonata medium with works for almost every known instrument, but as far as the form was concerned his innovations were of minor significance. Michael Tippett (born 1905) in his *Second Symphony* and sonatas (*e.g.*, for piano; for four horns) uses tonality in a fresh and valid way, and he has effected a stimulating rapprochement of the sonata form with the equal-voice polyphony characteristic of the English fantasia and madrigal (a genre of part-song) of Elizabethan and Jacobean times. The *Second Symphony* of Wilfred Josephs (born 1927) shows yet another potentially valuable reinterpretation of the fused-movement approach to the sonata: its long first movement serves the function of exposition, three intermediate movements act on one level as development and on another level as a combination of slow movement and scherzo, and a brief finale serves as a kind of recapitulation.

Other musical approaches use metre and instrumental tone colour to mark important musical points much as traditional sonata form used contrast of keys. Elliott Carter (born 1908) has combined a use of germinal motifs with a new rhythmic technique known as "metric modulation," a controlled change of metre foreshadowed in Brahms by such passages as the end of the *Second Piano Concerto*. Carter's *Sonata for Cello and Piano* is an example of this use of metre. Carter also uses the idea of sharply differentiating the musical subject matter given to the individual instruments of an ensemble—a resource found earlier in the *Second String Quartet* of Charles Ives (1874–1954). Some of the many styles of Igor Stravinsky (1882–1971), particularly after his late adoption of the twelve-tone approach, make ingenious use of germinal motifs; but his music really bases its structure on the juxtaposition of large blocks of distinct musical character, rather than on "development" in the sense traditionally associated with the sonata.

Music in the 1970s is too various in form, medium, esthetic attitude, and social function to allow any confident predictions. But all of these examples suggest that the sonata, and its special manifestation, the sonata form, can still provide composers with fertile areas of activity. As in the time of Haydn, Mozart, and Beethoven, success will continue to reward those who develop musical forms that grow naturally from the specific principles of composition used in their works, much as the sonata form grew out of the principle of contrast, conflict, and resolution of tonalities that characterized the sonatas, symphonies, and chamber music of the 18th and 19th centuries.

BIBLIOGRAPHY. E. BORREL, *La Sonate* (1951), a general modern study; W. MELLERS, "The Sonata Principle from c. 1750," in *Man and His Music* (1957), a description of stylistic trends that led from the Baroque to the Classical attitude; M. BUKOFZER, *Music in the Baroque Era, from Monteverdi to Bach* (1947), a general survey covering the period of the pre-Classical sonata; D.F. TOVEY, "Sonata Forms," in *The Forms of Music* (1956), a brilliant study, *Essays in Musical Analysis*, vol. 1–2 (1935), perceptive program notes on specific symphonic works, *Essays and Lectures on Music* (1949), important studies of Haydn, Beethoven, Schubert, and Brahms, and *A Companion to Beethoven's Pianoforte Sonatas* (1931); F.H. MARKS, *The Sonata: Its Form and Meaning in Piano Sonatas by Mozart* (n.d.); H. KELLER,"Wolfgang Amadeus Mozart," in R. SIMPSON (ed.), *The Symphony*, vol. 1 (1966), an imaginative and learned application of modern analytical principles to Mozart's symphonies; H.C. ROBBINS LANDON, *The Symphonies of Joseph Haydn* (1955), a monumental, detailed study of one of the central areas of sonata-form history.

(B.Ja.)

Sophists

The Sophists were certain Greek lecturers, writers, and teachers in the 5th and 4th centuries BC, most of whom travelled about the Greek-speaking world giving instruction in a wide range of subjects in return for fees.

HISTORY OF THE NAME

The term sophist (Greek *sophistes*) had earlier applications. It is sometimes said to have meant originally simply "clever" or "skilled man," but the list of those to whom Greek authors applied the term in its earlier sense makes it probable that it was rather more restricted in meaning. Seers, diviners, and poets predominate, and the earliest Sophists probably were the "sages" in early Greek societies. This would explain the subsequent application of the term to the Seven Wise Men (7th–6th century BC) who typified the highest early practical wisdom, and to Pre-Socratic philosophers generally. When Protagoras, in one of Plato's dialogues (*Protagoras,* 317 a–b) is made to say that, unlike others, he is willing to call himself a Sophist, he is using the term in its new sense of "professional teacher," but he wishes also to claim continuity with earlier sages as a teacher of wisdom. Plato and Aristotle altered the meaning again, however, when they claimed that professional teachers such as Protagoras were not seeking the truth but only victory in debate and were prepared to use dishonest means to achieve it. This produced the sense "captious or fallacious reasoner or quibbler," which has remained dominant to the present day. Finally, under the Roman Empire the term was applied to professors of rhetoric, to orators, and to prose writers generally, all of whom are sometimes regarded as constituting what is now called the Second Sophistic movement (see below).

THE 5TH-CENTURY SOPHISTS

The Sophistic movement

The names survive of nearly 30 Sophists properly so-called, of whom the most important were Protagoras, Gorgias, Antiphon, Prodicus, and Thrasymachus. Plato protested strongly that Socrates was in no sense a Sophist—he took no fees, and his devotion to the truth was beyond question. But from many points of view he is rightly regarded as a rather special member of the movement. The actual number of Sophists was clearly much larger than 30, and for about 70 years, until *c.* 380 BC, they were the sole source of higher education in the more advanced Greek cities. Thereafter, at least at Athens, they were largely replaced by the new philosophic schools, such as those of Plato and Isocrates. Plato's dialogue *Protagoras* describes something like a conference of Sophists at the house of Callias in Athens just before the Peloponnesian War. Antimoerus of Mende, described as one of the most distinguished of Protagoras' pupils, is there receiving professional instruction in order to become a Sophist (*Protagoras,* 315 a), and it is clear that this was already a normal way of entering the profession.

Most of the major Sophists were not Athenians, but they made Athens the centre for their activities, although travelling continuously. The importance of Athens was doubtless due in part to the greater freedom of speech prevailing there, in part to the patronage of wealthy men like Callias, and even to the positive encouragement of Pericles, who was said to have held long discussions with Sophists in his house. But primarily the Sophists congregated at Athens because they found there the greatest demand for what they had to offer, namely, instruction to young men, and the extent of this demand followed from the nature of the city's political life. Athens was a democracy, and although its limits were such that Thucydides could say it was governed by one man, Pericles, it nonetheless gave opportunities for a successful political career to citizens of the most diverse backgrounds, provided they could impress their audiences sufficiently in the council and the assembly. After Pericles' death this avenue became the highroad to political success.

The Sophistic teaching

The Sophists taught men how to speak and what arguments to use in public debate. A Sophistic education was increasingly sought after both by members of the oldest families and by aspiring newcomers without family backing. The changing pattern of Athenian society made merely traditional attitudes in many cases no longer adequate. Criticizing such attitudes and replacing them by rational arguments held special attraction for the young, and it explains the violent distaste which they aroused in traditionalists. Plato thought that much of the Sophistic attack upon traditional values was unfair and unjustified. But even he learned at least one thing from the Sophists—if the older values were to be defended, it must be by reasoned argument, not by appeals to tradition and unreflecting faith.

Seen from this point of view, the Sophistic movement was a valuable function of Athenian democracy in the 5th century BC. It offered an education designed to facilitate and promote success in public life. All of the Sophists appear to have provided a training in rhetoric and in the art of speaking, and the Sophistic movement, responsible for large advances in rhetorical theory, contributed greatly to the development of style in oratory. In modern times the view occasionally has been advanced that this was the Sophists' only concern. But the range of topics dealt with by the major Sophists makes this unlikely, and even if success in this direction was their ultimate aim, the means they used were surely as much indirect as direct, for the pupils were instructed not merely in the art of speaking, but in grammar; in the nature of virtue (*aretē*) and the bases of morality; in the history of society and the arts; in poetry, music, and mathematics; and also in astronomy and the physical sciences. Naturally the balance and emphasis differed from Sophist to Sophist, and some offered wider curricula than others. But this was an individual matter, and attempts by earlier historians of philosophy to divide the Sophistic movement into periods in which the nature of the instruction was altered are now seen to fail for lack of evidence. The 5th-century Sophists inaugurated a method of higher education that in range and method anticipated the modern humanistic approach inaugurated or revived during the Renaissance.

NATURE OF SOPHISTIC THOUGHT

A question still discussed is whether the Sophists in general had any real regard for truth or whether they taught their pupils that truth was unimportant compared with success in argument. Plato's hostile judgment on both counts is still frequently repeated without question. The Platonic writings make frequent reference to what Plato calls "eristic" (Greek *eristikos,* "fond of wrangling") and "antilogic"; the two often have been incorrectly treated as identical. Eristic, for Plato, consists in arguments aimed at victory rather than at truth. Antilogic involves the assignment to any argument of a counterargument that negates it, with the implication that both argument and counterargument are equally true. Antilogic in this sense was especially associated with Protagoras; but Plato, no doubt correctly, attributes its use to other Sophists as well. He regards the use of antilogic as essentially eristic, whether it be used to silence an opponent by making his position seem self-contradictory, or whether it be used mechanically to negate any proposition put forward in debate. He concludes that the widespread use of antilogic is evidence that Sophists had no real regard for the truth, which must itself be free from antilogic.

But Plato himself believed, for much or possibly all of his life, that the phenomenal world was essentially antilogical inasmuch as no statement about it could be made possessing a greater degree of truth than the contradictory of that statement. For example, if a man is tall in relation to one object, he will be short in relation to another object. In so characterizing the phenomenal world, Plato certainly did not wish to be called eristic—he regarded the application of antilogic to the description of the phenomenal world as an essential preliminary to the search for the truth residing in the Platonic Forms, which are themselves free from antilogic.

Seen in this perspective, the Sophistic use of antilogic must be judged less harshly. To the extent that it was used irresponsibly to secure success in debate it was eristic, and the temptation so to use it must often have arisen. But where it was invoked in the sincere belief that antilogic

elements were indeed involved, or where it was used for analyzing a complex situation in order to reveal its complexity, then antilogic was in no way inconsistent with devotion to truth. This raises the question to what extent the Sophists possessed any general view of the world or gave expression to any genuine philosophical views, whether original or derived. Ancient writers, influenced by Plato and Aristotle, seem to have excluded the Sophists, apart from Protagoras, for their schematized accounts of early Greek thinkers. Modern writers have frequently maintained that, whatever else they were, the Sophists were in no sense philosophers. Even those who acknowledge the philosophical interest of certain particular doctrines attributed to individual Sophists often tend to regard these as exceptions and claim that, inasmuch as the Sophists were not a school but only independent teachers and writers, as a class they were not philosophers. Two questions are involved: whether the Sophists held common intellectual doctrines and whether some or all of these could actually be termed philosophical.

Philosophy of the Sophists

Among moderns, Hegel was one of the first to reinsert the Sophists into the history of Greek philosophy. He did so within the framework of his own dialectic, in which every thesis invokes its own opposite, or antithesis; thus he treated the Sophists as representing the antithesis to the thesis of the group of philosophers known collectively as the Pre-Socratics. Pre-Socratics such as Thales, Heracleitus, and Parmenides sought the truth about the external world with a bold enthusiasm that produced a series of explanations, each claiming to be correct. None of these explanations of the physical world paid attention to the observer and each was driven to reject more and more of the phenomenal world itself as unreal. Finally, with the Eleatics, a 5th-century school at Elea in Italy that held that reality is a static one, of which Parmenides and Zeno are representatives, little or nothing of the phenomenal world was left as real. This trend in turn produced a growing distrust of the power of human beings to attain knowledge of the ultimate basis of natural phenomena. Philosophy had reached an impasse, and there was a danger of complete skepticism. Such an extreme position, according to Hegel's view, provoked the "antithesis" of the Sophistic movement, which rejected the "thesis" of the objectivists and concentrated attention upon man rather than upon nature. To Hegel, the Sophists were subjective Idealists, holding that reality is only minds and their contents, and so philosophy could move forward by turning its attention to the subjective element in knowing. Reflection upon the contrast between the thought of the Sophists and that of their predecessors produced the "syntheses" of Plato and Aristotle.

Whether any of the Sophists actually were subjective Idealists may be doubted. The conclusion depends in part on whether Protagoras held that phenomena had subjective existence only, or whether he thought that all things perceived had objective existence but were perceived differently according to the nature of the percipient and their relation to him—*i.e.*, whether he interpreted phenomena subjectively or relativistically. It is fairly clear, however, that the Sophists did concentrate very largely upon man and human society, upon questions of words in their relations to things, upon issues in the theory of knowledge, and upon the importance of the observer and the subjective element in reality and in the correct understanding of reality.

This emphasis helps to explain the philosophical hostility of Plato and Aristotle. Particularly in the eyes of Plato, anyone who looks for the truth in phenomena alone, whether he interprets it subjectively or relativistically, cannot hope to find it there; and his persistence in turning away from the right direction virtually amounts to a rejection of philosophy and of the search for truth. Many a subsequent thinker for whom metaphysics, or the investigation of the deepest nature of reality, was the crowning achievement of philosophy has felt with Plato that the Sophists were so antimetaphysical that they have no claim to rank as philosophers. But in a period when, for many philosophers, metaphysics is no longer the most important part of philosophy and is even for some no part at all, there is growing appreciation of a number of problems and doctrines recurring in the discussions of the Sophists in the 5th and 4th centuries BC. In the 18th and early 19th centuries the Sophists were considered charlatans. Their intellectual honesty was impugned, and their doctrines were blamed for weakening the moral fibre of Greece. The charge was based on two contentions, both correct: first, that many of the Sophists attacked the traditionally accepted moral code; and second, that they explored and even commended alternative approaches to morality that would condone or allow behaviour of a kind inadmissible under the stricter traditional code.

Moral views of the Sophists

Much less weight has been attached to these charges since about the mid-19th century. First, many of the attacks on the traditional morality were in the name of a new morality that claimed to be of greater validity. Attacks upon particular doctrines often claimed that accepted views should be abandoned as morally defective. Furthermore, even when socially disfavoured action seemed to be commended, this was frequently done to introduce a principle necessary in any satisfactory moral theory. Thus when Thrasymachus in the first book of Plato's *Republic* argues that justice is unwarranted when it merely contributes to another's good and not to the good of the doer, Plato agrees. Finally, there is no evidence that any of the Sophists were personally immoral or that any of their pupils were induced to immoral actions by Sophistic teaching. The serious discussion of moral problems and the theory of morality tends to improve behaviour, not to corrupt it.

WRITINGS

In addition to their teaching, the Sophists wrote many books, the titles of which are preserved by writers such as Diogenes Laërtius, who probably derived them from library catalogues. It has usually been supposed that the writings themselves hardly survived beyond the period of Plato and Aristotle, but this view requires modification in the light of papyrus finds, admittedly few, that were copied from Sophistic writings in the Christian Era. It also has been possible to identify in the works of later writers certain imitations or summaries of 5th-century Sophistic writers, whose names are unknown. The most important of these are the discussion of law in the *Protrepticus*, or "Exhortation to Philosophy," by the 3rd-century AD Syrian Neoplatonist Iamblichus, and the so-called *Dissoi logoi* found in the manuscripts of Sextus Empiricus (3rd century AD). This evidence suggests that while most later writers took their accounts of the Sophists from earlier writers, especially from Plato, the original writings did in many cases survive and were consulted.

PARTICULAR DOCTRINES

As part of his defense of the Sophists against the charge of immoral teachings, the English historian George Grote (1794–1871) maintained that they had nothing in common with each other except their profession, as paid teachers qualifying young men to think, speak, and act with credit to themselves as citizens. This denial of common doctrines cannot be sustained—the evidence is against it. While the Sophists were not a sect, with a set of obligatory beliefs or doctrines, they had a common interest in a whole series of questions to which they sought to apply solutions along certain clearly defined lines.

There are great difficulties, however, in the precise reconstruction of individual Sophistic doctrines. No complete writings survive from any of the Sophists to check the accounts found in Plato, and later writers were often, but not always, dependent upon what they found in Plato. Plato doubtless knew well the doctrines of individual Sophists; but he was writing for those to whom these doctrines were already well known, and he was always more interested in following the argument where it led than in providing precise statements of other people's views for the sake of posterity. Consequently, almost everything that is said about particular Sophistic doctrines is subject to controversy.

Theoretical issues. Relativism and skepticism have often been regarded as common features of the Sophistic

movement as a whole. But it was early pointed out that only in Protagoras and Gorgias is there any suggestion of a radical skepticism about the possibility of knowledge; and even in their case Sextus Empiricus, in his discussion of skepticism, is probably right when he declares that neither was really a skeptic. Protagoras does seem to have restricted knowledge to sense experience, but he believed emphatically that whatever was perceived by the senses was certainly true. This led him to assert that the tangent does not touch the circle at a point only, but along a definite length of the circumference; clearly he was referring to human perception of drawn tangents and circles. Gorgias, who claimed that nothing exists, or if it does exist it cannot be known, or if it exists and is knowable it cannot be communicated to another, has often been accused of denying all reality and all knowledge. Yet he also seems to have appealed in his very discussion of these themes to the certainty of perceived facts about the physical world; *e.g.*, that chariots do not race across the sea. Others dismiss his whole thesis as a satire or joke against philosophers.

Relativism and skepticism

Probably neither view is correct. What Gorgias seems to have been attacking was not perceived reality nor one's power to perceive it but the attempt to assign existence or nonexistence (with the metaphysical implications of such an operation) to what we perceive around us. There is evidence that other Sophists (*e.g.*, Hippias) were interested in questions of this kind, and it is likely that they were all concerned to some degree with rejecting claims of any nonsensible existence, such as those of the Eleatics. The Sophists, in fact, were attempting to explain the phenomenal world without appealing to any principles outside of phenomena. They believed that this could be done by including the observer within the phenomenal world. Their refusal to go beyond phenomena was, for Plato, the great weakness in their thinking.

A second common generalization about the Sophists has been that they represent a revolt against science and the study of the physical world. The evidence is against this, inasmuch as for Hippias, Prodicus, Gorgias, and Protagoras there are records of a definite interest in questions of this kind. The truth is rather that they were in revolt against attempts to explain the physical world by appeals to principles that could not be perceived by the senses; and instead of framing new "objective" explanations, they attempted to explain things, where explanation was required, by introducing the perceiver as one element in the perceptual situation.

Nature and law

One of the most famous doctrines associated with the Sophistic movement was the opposition between nature and custom or convention in morals. It is probable that the antithesis did not originate in Sophistic circles but was rather earlier; but it was clearly very popular and figured largely in Sophistic discussions. The commonest form of the doctrine involved an appeal from conventional laws to supposedly higher laws based on nature. Sometimes these higher laws were invoked to remedy defects in actual laws and to impose more stringent obligations; but usually it was in order to free men from restrictions unjustifiably imposed by human laws that the appeal to nature was made. In its extreme form the appeal involved the throwing off of all restraints upon self-interest and the desires of the individual (*e.g.*, the doctrine of Callicles in Plato's *Gorgias* that might, if one possesses it, is actually right), and it was this, more than anything else, that gave support to charges against the Sophists of immoral teaching. On other occasions the terms of the antithesis were reversed and human laws were explicitly acclaimed as superior to the laws of nature and as representing progress achieved by human endeavour. In all cases the laws of nature were regarded not as generalized descriptions of what actually happens in the natural world (and so not like the laws of physics to which no exceptions are possible) but rather as norms that people ought to follow but are free to ignore. Thus the appeal to nature tended to mean an appeal to the nature of man treated as a source for norms of conduct.

To Greeks this appeal was not very novel. It represented a conscious probing and exploration into an area wherein, according to their whole tradition of thought, lay the true source for norms of conduct. If Callicles in Plato's *Gorgias* represents a position actually held by a living Sophist when he advocates free rein for the passions, then it was easy for Plato to argue in reply that the nature of man, if it is to be fulfilled, requires organization and restraint in the license given to the desires of particular aspects of it; otherwise the interests of the whole will be frustrated. Both Plato and Aristotle, in basing so much of their ethics on the nature of man, are only following up the approach begun by the Sophists.

Religion and society

Humanistic issues. The Sophists have sometimes been characterized by their attacks on the traditional religious beliefs of the Greeks. It is true that more than one Sophist seems to have faced prosecution for impiety, as did Socrates also. Protagoras wrote "concerning the gods, I cannot know either that they exist or that they do not exist nor what they are like in form," and Prodicus offered a sociological account of the development of religion. Critias went further when he supposed that the gods were deliberately invented to inspire fear in the evildoer. It is thus probably correct to say that the tendency of much Sophistic thought was to reject the traditional doctrines about the gods. Indeed this follows almost inevitably if the supposition is correct that all the Sophists were attempting to explain the phenomenal world from within itself, while excluding all principles or entities not discernible in phenomena. But in their agnostic attitudes toward the Olympian deities the Sophists were probably at one with most of the Pre-Socratic philosophers of the 6th and 5th centuries and also with most thinking people living toward the end of the 5th century. It is thus probably misleading to regard them as revolutionary in their religious beliefs.

The importance the Sophists attached to man meant that they were extremely interested in the history and organization of human societies. Here again most is known about Protagoras, and there is a danger of treating his particular doctrines as typical of the Sophistic movement as a whole. In the 5th century, human history was very commonly seen in terms of a decline from an earlier golden age. Another view supposed that there were recurring cycles in human affairs according to which a progression from good to bad would give way to one from bad to good. The typical Sophistic attitude toward society rejected both of these views in favour of one that saw human history in terms of progress from savagery to civilization. In a famous myth Protagoras explained how man achieved civilized society first with the aid of arts and crafts and then by gaining a sense of respect and justice in the ordering of his affairs. The general thinking of most of the Sophists seems to have been along similar lines.

One of the most distinctive Sophistic tenets was that virtue can be taught, a position springing naturally from the Sophists' professional claim to be the teachers of young men. But the word virtue (*aretē*) implied both success in living and the qualities necessary for achieving such success, and the claim that *aretē* could be taught by the kind of teaching that the Sophists offered had far-ranging implications. It involved the rejection of the view that *aretē* came only by birth—for example, by being born a member of a noble family—and it involved also the rejection of the doctrine that *aretē* was a matter of the chance occurrence of specified qualities in particular individuals. *Aretē,* in the Sophists' view, was the result of known and controllable procedures, a contention of profound importance for the organization of society. Moreover, what can be taught has some relation to what can be known and understood. The belief that teaching of a high intellectual calibre could produce success both for the individual and for governments has had a profound influence upon the subsequent history of education. Once again, it is through the acceptance of this doctrine by Plato and Aristotle that the Sophistic position came to be part of subsequent humanist tradition.

THE SECOND SOPHISTIC MOVEMENT

It is a historical accident that the name "Sophist" came to be applied to the Second Sophistic movement. Greek liter-

ature underwent a period of eclipse during the 1st century BC and under the early Roman Empire. But Roman dominance did not prevent a growing interest in sophistic oratory in the Greek-speaking world during the 1st century AD. This oratory aimed merely at instructing or interesting an audience and had of necessity no political function. But it was based on elaborate rules and required a thorough knowledge of the poets and prose writers of antiquity. Training was provided by professional teachers of rhetoric who claimed the title of Sophists, just as the 5th-century Sophists had adopted a name already used by others.

The revival of the Greek spirit under Hadrian and other emperors in the 2nd century AD who were also admirers of Greek culture found expression in a fresh flowering of Greek prose following principles developed and applied by the professors of rhetoric in the 1st century AD. Hence a group of Greek prose writers in the 2nd century AD were regarded as constituting the Second Sophistic movement. This was a backward-looking movement that took as its models Athenian writers of the 5th and 4th centuries BC; hence the label "Atticists" (Greek *Attikos*, "Athenian") applied to some of its leading members. The limits of the movement were never clear. It is usually taken to include Polemon of Athens, Herodes Atticus, Aelius Aristides, Maximus of Tyre, and the group of Philostrati. Dio Chrysostom of Prusa is often included, although others would regard him as preparing the way for the main period. Other writers, like Lucian, Aelian, and Alciphron, were influenced by the movement even if not properly members of it; and the writers of prose romances, such as Longus and Heliodorus, and the historians Dio Cassius and Herodian are also associated with the general trend. By the 3rd century AD, however, its impulse was weakening, and was shortly no longer distinguishable within the general stream of Greek literature.

BIBLIOGRAPHY

Fifth-century Sophists (Ancient sources and fragments): R.K. SPRAGUE, *The Older Sophists, a Complete Translation* (1972), and in K. FREEMAN, *Ancilla to the Pre-Socratic Philosophers* (1948), to be used together with Freeman's *The Pre-Socratic Philosophers*, 2nd ed. (1949). The original Greek texts are in H. DIELS and W. KRANZ (eds.), *Die Fragmente der Vorsokratiker*, 6th ed., vol. 2 (1952); and are edited with an Italian translation and commentary in M. UNTERSTEINER, *Sofisti*, 4 vol. (1949–62). *(General discussions):* T. GOMPERZ, *Greek Thinkers*, Eng. trans. by L. MAGNUS, vol. 1, bk. 3, ch. 5–7 (1901); M. UNTERSTEINER, *The Sophists*, Eng. trans. by K. FREEMAN (1954); W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 3, *The Fifth-Century Enlightenment* (1969).

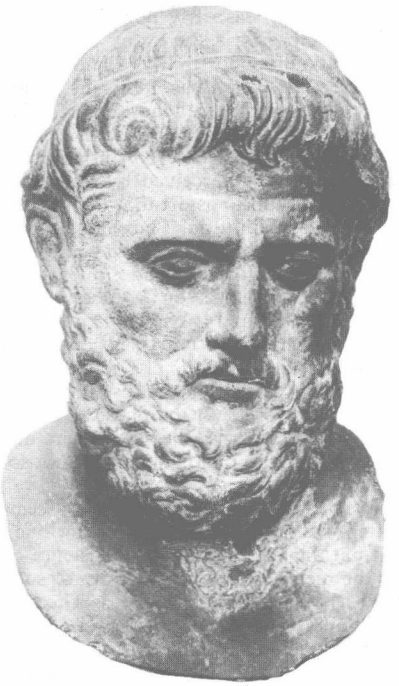
Second Sophistic period: G.W. BOWERSOCK, *Greek Sophists in the Roman Empire* (1969).

(G.B.K.)

Sophocles

More than any other dramatist of classical Greece—including his great fellow tragedians Aeschylus and Euripides—the poet Sophocles gave profound form to complex dilemmas of experience in the 5th century BC, while his native Athens grew to its fullest cultural, political, and economic development. His plays not only earned him the highest respect throughout his long life but they also translate the essential elements of classical civilization into permanently gripping theatre.

Public life. Sophocles was born *c.* 496 BC, when the Athenians had been experimenting for only a dozen years with a new, limited democratic rule. By the time he died, in 406, Athens was about to surrender to a force of Greek states (led by Sparta) that had banded together to overthrow the city's commercial and imperial "tyranny"; its democracy, too, seemed in danger of collapsing. Yet in his last play, *Oedipus at Colonus,* Sophocles sang the praises both of his own birthplace, Colonus, a village outside the walls of Athens, and of the great city itself—as though it were still in its heyday of 431, before the long war against Sparta had begun. Sophocles' faith in the city and its ideals—even during its troubles and defeats—bears out an ancient report that he refused invitations from kings to settle abroad at their courts.



Sophocles, bronze bust copied from a Greek original, 340–330 BC. In the Museo Archeologico, Florence.
Alinari

An event of Sophocles' adolescence could be said to have set the pattern for his future. Because of his beauty of physique, his athletic prowess, and his skill in music, he was chosen in 480, when he was 16, to lead the paean (formal choral chant dedicated to a god) celebrating a decisive Greek sea victory over the Persians at Salamis. At this turning point of Athenian history, which inaugurated 50 years of security, immense enterprise, maritime expansion, and cultural achievement, Sophocles was already a popular favourite, participating actively in the community and in religious ceremony and exercising outstanding artistic talents.

Public service

All the relatively meagre information about Sophocles' civic life is consistent with this. In 442 he served as one of the treasurers responsible for receiving and managing tribute money from some 300 subject states of the Athenian Empire. In 440 he was elected one of ten generals (high executive officials, as well as commanders of the armed forces); and, as a colleague of Pericles (died 429), the greatest leader of the period, he took part in an expedition to bring a wavering "ally" back into line. He served again as general perhaps two other times later. Around 421, while acting as priest for a religious society honouring a minor god of healing, he kept a snake sacred to the divine physician Asclepius in his own house until a temple under construction was ready to receive it. In 413, then aged 83, Sophocles was one of ten advisory commissioners granted special powers and entrusted with organizing Athens' financial and domestic recovery after a terrible defeat at Syracuse in Sicily.

These few facts are about all that is known of Sophocles' life. They imply distinguished, steady, earnest attachment to Athens, its government, religion, and social forms. Some scholars have inferred that Sophocles remained untroubled by the issues inherent in the growth of Athenian power, the party conflicts, and the cultural upheaval of the 5th century; and, drawing on the remarks of ancient writers, they have called him "genial" and "serene" in temperament and "orthodox" (if not reactionary) in viewpoint. If the tragedies had not survived, this conclusion might be acceptable. For Sophocles was a gentleman of the leisure class, wealthy by birth (his father probably a manufacturer of armour), highly educated, socially cultivated, noted for his grace and charm, on easy terms with aristocratic families, a personal friend of men such as Cimon, the conservative statesman, and He-