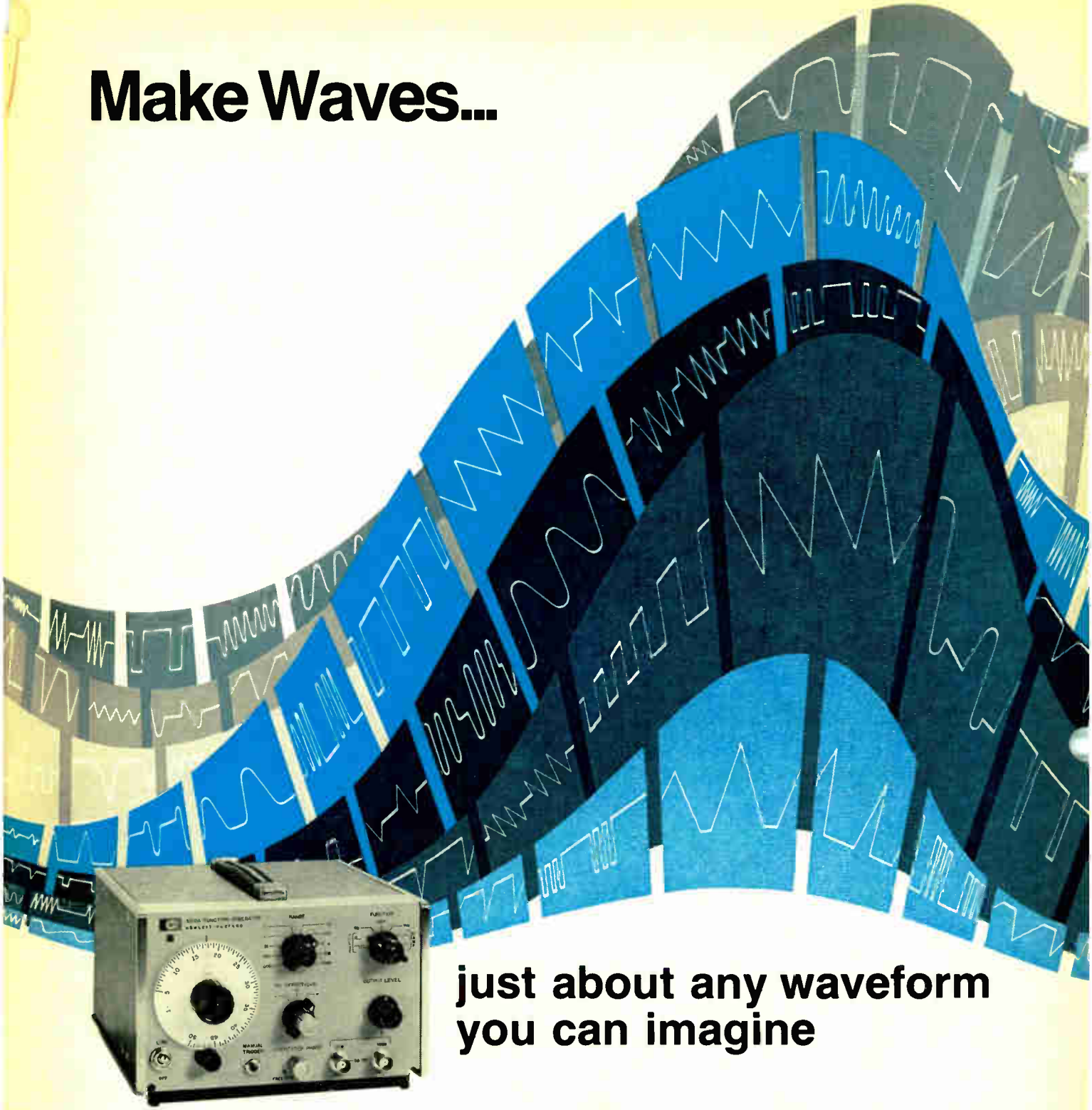
IEEE spectrum

features

Make Waves...



just about any waveform you can imagine

HP's 3310A is the function generator that gives you seven different waveforms—in three different modes—in one inexpensive package.

In its basic form, the 3310A gives you a continuous output of square waves, sine waves, and triangle waves – plus positive and negative ramps and pulses—for only $595.

By adding HP's new Option H10 (only $140), you can generate each of these seven waveforms in two other modes—single-cycle and multiple cycle "bursts." These "bursts" can be triggered either manually or by an external oscillator; starting-point phase can be varied by ± 90°.

With or without Option H10, the 3310 gives you a choice of ten frequency ranges—from 0.0005 Hz to 5 MHz—and an output voltage range from 15 mV pk-pk to 15 V pk-pk into 50Ω load. Dc offset of ±5 V into 50Ω load is also standard.

With Option H10, the 3310A can be used in frequency-response and transient-response testing, as a waveform converter, for generating phase-coherent waveforms, and as a frequency multiplier or divider, among other things. Applications include testing television and communications systems, radar systems, and analog or digital circuits.

For further information on the 3310A and Option H10, contact your local HP field engineer, or write to Hewlett-Packard, Palo Alto, California 94304. In Europe: 1217 Meyrin-Geneva, Switzerland.

090/41

HEWLETT hp PACKARD

SIGNAL SOURCES

Circle No. 4 on Reader Service Card

the cover

Despite the fact that electronic music synthesis has become an indispensable part of today's recording industry, most classical music is recorded without such electronic modifications. Whether or not classics like Mozart's Piano Concerto No. 20 in D Minor (cover) would be enhanced by these techniques is left to the judgment of the reader, who is invited to examine what it's all about in the article beginning on page 18

departments

spectrum

David DeWitt, IBM Corporation, *Editor*

EDITORIAL STAFF Ronald K. Jurgen, *Managing Editor;* Robert E. Whitlock, *Senior Editor;* Gordon D. Friedlander, *Senior Staff Writer;* Marcelino Eleccion, *Staff Writer;* Evelyn Tucker, Alexander A. McKenzie, Charles W. Beardsley, *Associate Editors;* Pamela L. Abraham, *Assistant Editor;* Stella Grazda, *Editorial Assistant;* Ruth M. Edmiston, *Production Editor;* Herbert Taylor, *Art Director;* Janet Mannheimer, *Assistant Art Director;* Morris Khan, *Staff Artist*

PUBLICATIONS OPERATIONS Elwood K. Gannett, *Director, Editorial Services;* Alexander A. McKenzie, *Assistant to the Director;* Patricia Penick, *Administrative Assistant to the Director*

W. J. Hilty, *Director, Convention and Publishing Services;* William R. Saunders, *Advertising Director for Publications;* Henry Prins, *Manager, Marketing Services and Research;* Carl Maier, *Advertising Production Manager;* Oksana Bryn, *Assistant Advertising Production Manager*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE SPECTRUM is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters address: 345 East 47 Street, New York, N.Y. 10017. Cable address: ITRIPLEE. Telephone: 212—752-6800. Published at 20th and Northampton Sts., Easton, Pa. 18042. **Change of address** must be received by the first of a month to be effective for the following month's issue. Please send the SPECTRUM mailing label showing your old address, together with your new address, to Coding Department, IEEE, 345 E. 47 St., New York, N.Y. 10017. **Annual subscription:** IEEE members, first subscription $3.00 included in dues. Single copies $1.50. Nonmember subscriptions and additional member subscriptions available in either microfiche or printed form. Prices obtainable on request. **Editorial correspondence** should be addressed to IEEE SPECTRUM at IEEE Headquarters. **Advertising correspondence** should be addressed to IEEE Advertising Department, at IEEE Headquarters. Telephone: 212—752-6800.

OTHER IEEE PUBLICATIONS: IEEE also publishes the PROCEEDINGS OF THE IEEE and more than 30 Transactions for IEEE Groups with specialized interests within the electrical and electronics field. Manuscripts for any IEEE publication should be sent to the editor of that publication whose name and address are shown on page 93 of the January issue. When in doubt, send the manuscript to E. K. Gannett, Editorial Services, at IEEE Headquarters, for forwarding to the correct party.

IEEE also copublishes, with the Institution of Electrical Engineers of London, England, *Electrical & Electronics Abstracts* and *Computer & Control Abstracts*, as well as their companion titles journals, *Current Papers in Electrical & Electronics Engineering* and *Current Papers on Computers and Control.*

Spectral lines

Technology and the social system. What is technology's place in the social system? Some think there is too much of it; others feel it is out of control; still others argue that its uninhibited progress not only should not be threatened but should be accelerated, and that in any case nothing can stop it; and to many professionals only technology can save mankind from technology. A substantial number, possibly already a majority in the United States, feel that it is driving us toward various catastrophes, including undermining the concept of democracy, unless we . . . But confusion reigns as to just what to do.

Out of this has developed two schools of thought—the doomsday school and the we'll-find-a-way school. A surprised society has awakened to the fact that just devising methods and things that we think are useful is not the same as producing and applying them in ways that are truly beneficial. We are beginning to see that we must reverse the two-century-old trend of our social mores adapting themselves to technology and devote our attention to having technology serve society.

There is a clear call today for engineers to view and understand their true relation to the outside world, and from this broader viewpoint, to see objectively and imaginitively how their special capabilities can be applied effectively to benefit mankind. To my mind technologists are exceptionally well placed and have unusual assets for the as-yet-undefined tasks ahead. Will engineers as a group respond to society's call or continue as though the future is likely to pattern the past?

Between technology and society lies the policy maker. If technologists wish to influence the development and application of technology, they must reach the policy maker, for it is he who interprets what society should have and instructs technologists what to do. There are two approaches to him: the direct one is to provide reliable, unbiased, easy-to-use data, analyses, and uncertainties to help him reach rational decisions; the indirect one is to provide the public and all interested parties with the same high-quality information on national and local issues where technology is involved, as a foundation to rational public debate.

The direct approach to the policy maker has received increasing attention in the last few years. We are still a long way, however, from top policy makers accepting the techniques and being seriously influenced by them. Technologists can make important contributions to this area. In my opinion, it is the most important technical area for the nation, one that the educational establishment should develop, both through research and by the development of talent that will soon be needed. Without such analytical ability we have little chance of making reasonably correct long-range decisions.

In the indirect approach via the public, technologists have the enviable status of an unequaled reputation for objectivity and freedom from partisan bias, and an unmatched recognition of ability to perform. (If they can send men to the moon and back, what can they not do, if only they put their minds to it?) This reputation is a thing of great value to be protected, nurtured, and developed.

There are other broad doctrines of behavior that will not be discussed here. Individually, engineers contribute by accepting these doctrines; but as a group much more can be achieved. The IEEE is one such group. In the past it has limited itself to being the medium through which the members of its component disciplines support each other by the exchange of technical information. To reach closer to the basic problem of assisting its disciplines to truly benefit society, however, it must consider expanding its responsibilities to include, to a far greater extent than it has historically, the flow of information across its interfaces with the outside world.

There is need for a center of unquestionable standing, separated from efforts to improve the economic or social status of its members, but dedicated rather to clarifying components of problems involving its area of technology, out lining what is known about them, what is not known, and what might be ascertained, and uncovering specific differences of a controversy. It would make no recommendations, but the information it will provide will tend to channel debates more rationally than is commonly the case today.

I can visualize a panel made up of members of various disciplines, some nontechnical, thus providing a wide range of viewpoints, taking responsibility for clarifying to the public selected important problems.

In view of our changing world, of the changing status of technology, IEEE must review its responsibilities to society and help bring its disciplines closer to its basic problems. Society can reasonably expect at least three things of a group of technologists: (1) to break down the isolation of disciplines; (2) to move the technical disciplines closer to the real problems of society; (3) to develop technical centers of unimpeachable quality to which all can turn for reliable, unbiased information.

Raymond M. Wilmotte

The IEEE has started to increase its service to the policy makers. Next month in this space, Harold Chestnut, Vice President, Technical Activities, will tell what is being done and what method must still be devised to make us more effective.

David DeWitt, Editor

Electronic music synthesis for recordings

Employed in the production of practically every form of recorded music today, electronic music synthesis in some way affects 95 percent of the multibillion-dollar recording industry

Harry F. Olson RCA Laboratories

Electronic music synthesis is a process whereby several parts or elements of a musical composition that are performed or produced as separate entities are combined to form the entire or rendered composition. Since the entire process can seldom be carried out in real time, the synthesis of music has not been a performing-type rendition; therefore, the final product takes the form of a record that can be reproduced at any given time. Electronic music synthesis includes the modification and combination of conventional and original sound sources, manual and programmed electronic music synthesizers, and digital computers.

To dispel the confusion that sometimes arises with respect to the definition of electronic music synthesis, the author would like to state that the term is not synonymous with music that is produced by an electronic music synthesizer. In the process of electronic music synthesis, the individual elements of a particular music score are performed separately by independent means and later combined electronically to form the complete composition. With this definition in mind, it should come as no surprise that almost all popular and contemporary ("rock") music that is recorded today is in fact electronically synthesized.

Properties of a musical tone[1,2]

The ultimate objective of electronic music synthesis is the production of desired musical tones. In electronic music synthesis, whether the process involves modification, combination of existing tones, or the production of tones by manual or programmed synthesis, the foundation for all synthesis involves the basic physical properties of a musical tone—frequency, intensity, waveform, and time. A more meaningful and useful description of a musical tone can be given in terms of frequency, intensity, growth, steady state, decay, duration, portamento, timbre (spectrum), vibrato, and deviation. These properties of a musical tone are depicted in Fig. 1.

The following definitions of tone properties are descriptive rather than absolutely rigorous presentations of the formal and somewhat abstruse language of the standards.

Frequency of a sound wave is the number of cycles occurring per unit of time, measured in hertz. The subjective counterpart of frequency is pitch.

Intensity of a sound wave is energy transmitted per unit of time. The intensity is usually expressed in decibels above the threshold of hearing at 1000 Hz, which is 0 dB = 10^{-16} W/cm^2. The subjective counterpart of intensity is loudness.

Growth is the time required for a sound to build up to some fraction of the ultimate value.

Steady state of a sound is the length of time in which there is no change in the intensity.

Decay is the time required for sound to fall to some fraction of the original value. Note that growth, steady state, and decay lumped together become the envelope of a tone.

Duration is the length of time that a sound persists without interruption or discontinuity in the output.

Portamento is a uniform glide in frequency from a sound of one frequency to a sound of another frequency.

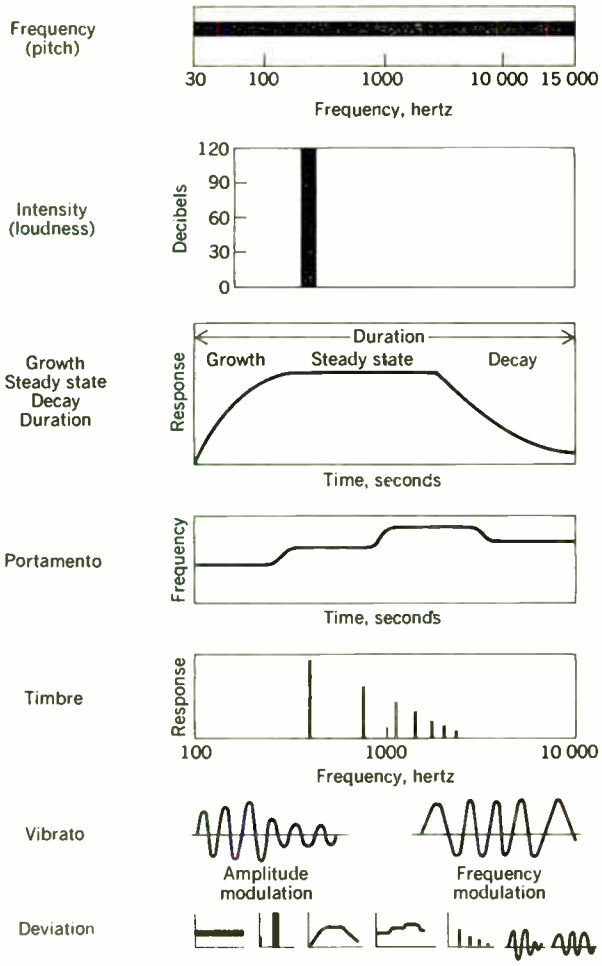


FIGURE 1. Properties of a musical tone.

Portamento is also termed a frequency glide.

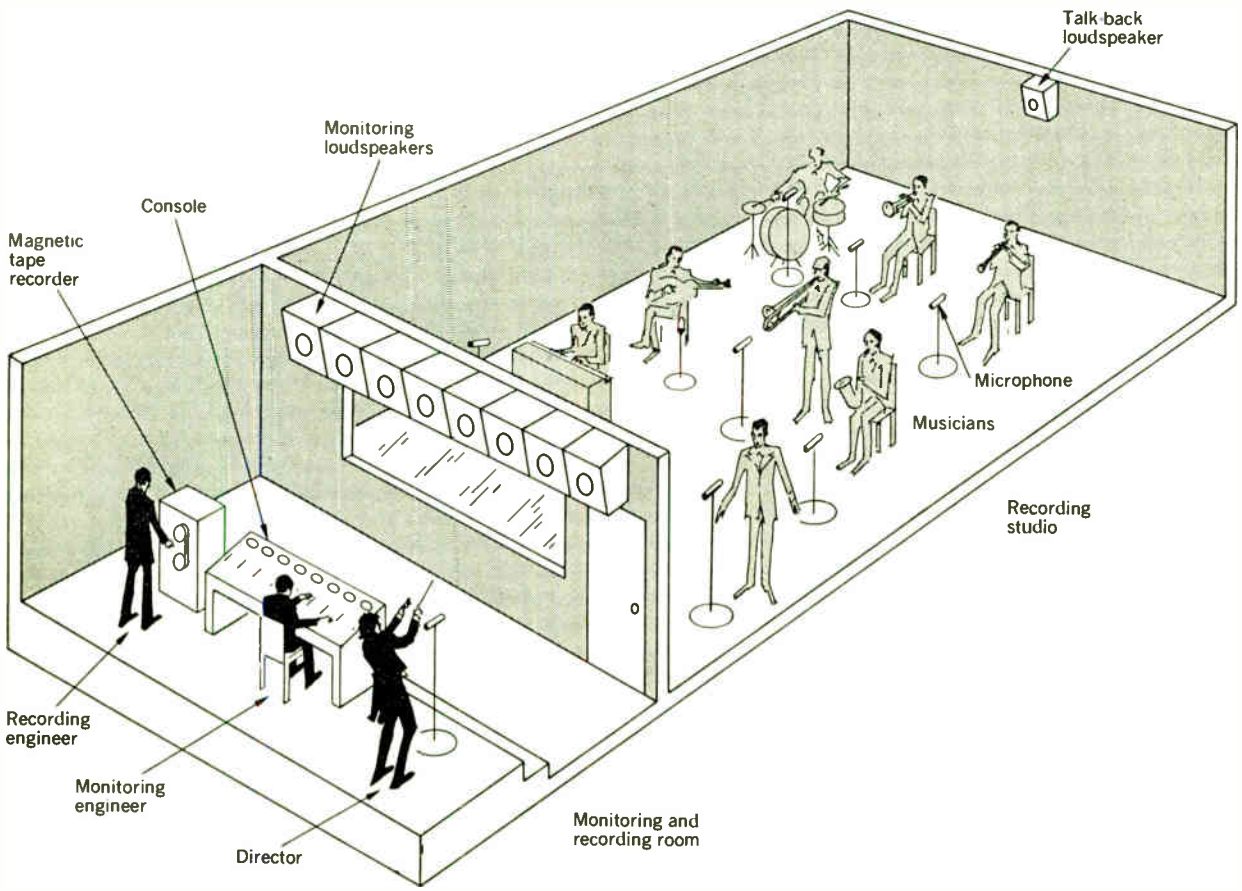
A complex sound wave is made up of the fundamental tone and overtones. The *timbre* or spectrum of a tone is expressed in the number, intensity, and phase relations of the components; that is, the fundamental and overtones or partials.

Vibrato is a low-frequency modulation of a musical tone. This may result from either frequency modulation or amplitude modulation or a combination of both. In general, the modulation frequency is of the order of 7 Hz. *Tremolo*, a special case of vibrato, is created by amplitude modulation only.

Deviation is a departure from the regular and is one of the beautiful and artistic characteristics of some types of music.

With reference to the preceding descriptions, many of the properties of a tone are interdependent. For example, timbre is influenced by the attack, decay, portamento, vibrato, etc. When the properties of a tone as just defined and depicted in Fig. 1 are specified, the tone can be completely described. Furthermore, the tone can be produced from these specifications by providing electronic means for generating its characteristic properties.

FIGURE 2. Perspective view of a recording studio depicting the sound pickup of eight musicians by means of separate microphones and the monitoring room with director and recording engineers operating the console and recorder.



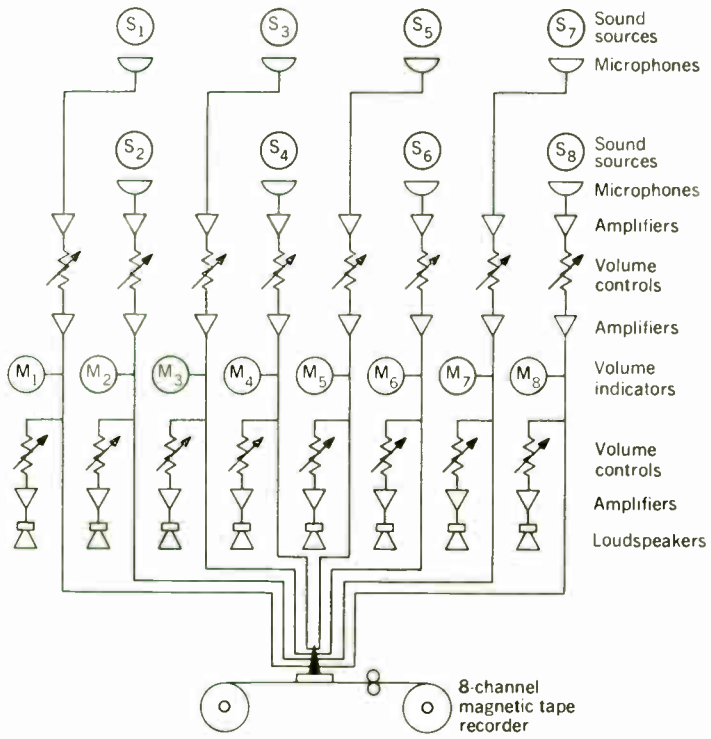


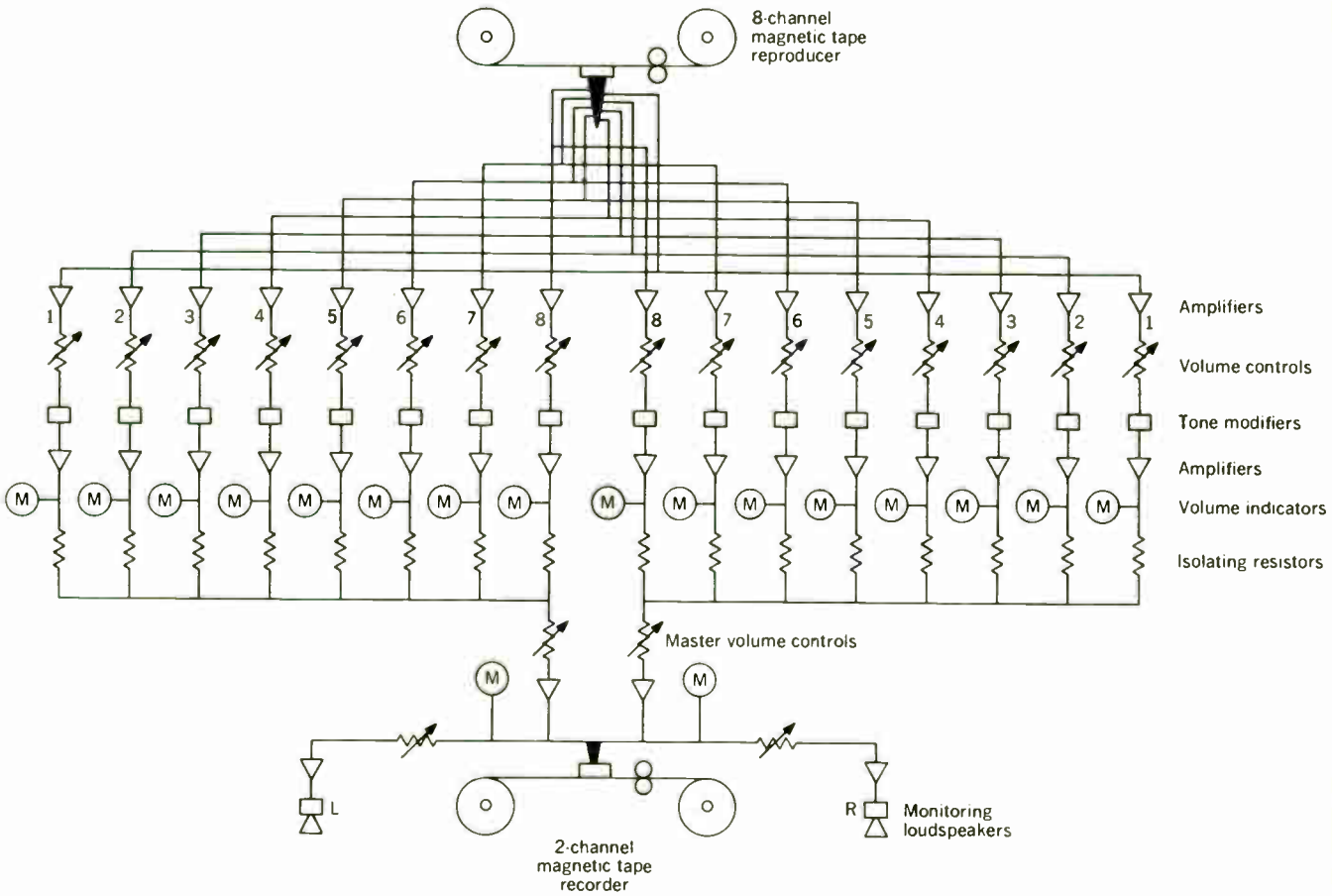
FIGURE 3. Schematic diagram of the recording of eight sound sources, each with a pickup microphone enabling the outputs of the sound sources to be individually recorded on eight separate tracks of a tape recorder.

Electronic music synthesis from conventional sound sources[3-5]

Since the introduction of magnetic tape recording and until only a few years ago, master tape recordings were recorded on two- and three-channel machines. The recording of original sounds on 8-, 16-, and 24-channel machines is a recent innovation brought about by the requirements of new sounds in popular and contemporary music in order to produce commercially successful records. The use of the 8-, 16-, and 24-channel magnetic tape recorder makes it possible to record each instrument or performer on a separate channel as depicted by the perspective view of the studio, monitoring room, musicians, conductor, and engineers in Fig. 2. A schematic diagram of the elements used in the recording system is shown in Fig. 3.

The eight-channel system is used in this article for the sake of simplicity. In order to record each and every sound source, that is, record each instrument or performer separately when playing as a group, there should be a minimal acoustic crosstalk between the different sound sources. To reduce the crosstalk, either bidirectional or unidirectional microphones are employed. In addition, a studio with a very low reverberation time is used, and each sound source is fed to a separate track so that the isolation is preserved. In turn, each sound source is monitored by means of volume indicators and loudspeakers. The use of a loudspeaker in each channel makes it possible to detect acoustic crosstalk and other undesirable effects.

FIGURE 4. The mixing, modifying, and conversion to two-channel stereophonic magnetic tape from the eight sound sources recorded in Figs. 2 and 3.



Recording each sound source on a separate track represents the ultimate for electronic music synthesis. In some performances, the recording of two or more sound sources on one track may be desirable from the standpoint of simplifying the procedure of electronic music synthesis. In any event, the synthesis procedure is essentially the same.

A schematic diagram for electronic synthesizing of the recording produced by the system of Figs. 2 and 3 is shown in Fig. 4. In this particular case, the final product is a two-channel stereophonic record. The system of Fig. 4 provides a facility for placing any of the eight sound sources in either the left or right channel or any mixture in both channels. The important element in this system is the tone modifier, which is diagramed in Fig. 5. There are 16 of these tone modifiers in the system of Fig. 4.

Signal delay in combination with signal level makes it possible to specify the auditory perspective of any sound source, that is, place any sound source in reproduction at either the right or left loudspeaker or any position between. The control of the auditory perspective is illustrated in Fig. 6. If the entire audio signal is fed to the left loudspeaker and no audio signal to the right loudspeaker, the sound source will appear to be located at position 1. In the same way, if the entire signal is fed to the right loudspeaker and no signal to the left loudspeaker, the sound source will appear to be located at position 5. If the same audio signal is fed to both loudspeakers, the sound source will appear to be located at position 3. However, if delay is introduced in the left channel but the same amplitude is fed to both loudspeakers, the sound source will be moved from location 3 (no delay) to anywhere between 3 and 5, depending upon the amount of delay. (A large delay will produce location 5.) If the delay in the two channels is the same but the signal in the left channel is attenuated, then the sound source can be moved from location 3 (no attenuation) to anywhere between 3 and 5, depending upon the amount of attenuation. (Location 5 results from considerable attenuation.) Employing both delay and attenuation, therefore, places the sound source in any position at or between the loudspeakers. The attenuation and delay provide a powerful means for producing the desired auditory perspective.

Let us return now to the tone modifier of Fig. 5: The frequency-response modifier consists of frequency-selective networks that change the original timbre (waveform) by accentuating or attenuating frequency bands.

The timbre modifier of Fig. 5 is somewhat similar to the frequency-response modifier save that new frequency components may be added.

The vibrato or tremolo generator modulates, by frequency or amplitude modulation or a combination of both, the original audio signal, with a modulation frequency of the order of 7 Hz.

Since the original sound as depicted in Figs. 2 and 3 is recorded without any reverberation, a means must be provided to introduce the effect artificially.

In general, the goal of all sound reproduction is to reduce nonlinear distortion. However, in some instances the subjective response can be heightened by the introduction of an appropriate nonlinear distortion. Accordingly, means are provided for introducing such nonlinear distortion.

The fuzz producer is a method of generating high harmonics and frequency-modified random noise. Here again is another example of the introduction of what is normally considered undesirable—noise—to heighten the artistic

FIGURE 5. An electronic modifier.

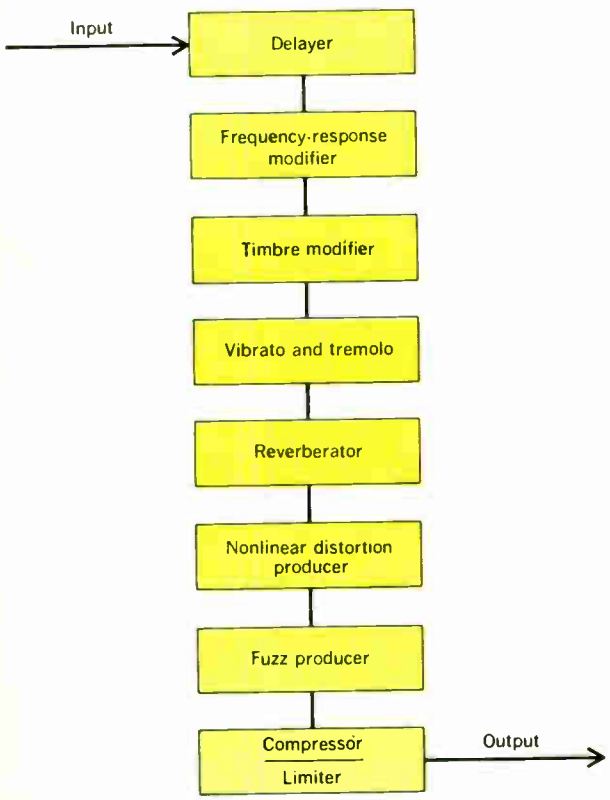
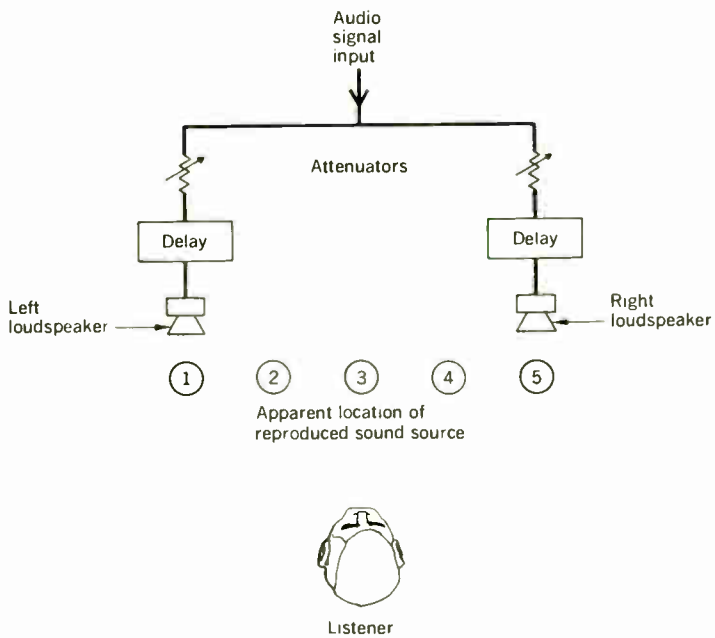


FIGURE 6. System of delayers and attenuators for placing the apparent sound of an audio signal at any point between two stereophonic loudspeakers.



aspects of the recorded sound.

The volume compressor of Fig. 5 is an electronic system that reduces the amplification in a gradual manner whenever the signal input attains a certain level. Compressors are often used to reduce the volume or amplitude range in such systems as sound motion pictures, magnetic tape and phonograph disk recording, sound broadcasting, and sound reinforcing. A reduction in the volume range of radio, magnetic tape, and phonograph sound reproduction makes it possible to reproduce the wide amplitude range of orchestra music in the home without excessive top levels. The use of the compressor improves the signal-to-noise ratio, thereby improving the intelligibility of speech and enhancing music sound reproduction when the ambient noise is high.

A limiter is similar to the compressor except that the relationship between the input and output is constant up to a certain input level. Beyond that level, the output does not increase. The limiter is extremely useful for protection against sudden overloads in the sound system.

An extension of the tone modifier is shown in Fig. 7. This particular modifier makes it possible to extend the frequency range of an instrument and modify its timbre. The fundamental selector searches and selects the fundamental frequency of the instrument. This fundamental frequency is then multiplied or divided to extend the frequency range of the instrument or to produce more than one instrument (as in Fig. 7). Under these conditions, the original sound source or instrument supplies the fundamental frequency (pitch), the envelope, and the intensity (loudness).

The systems of Figs. 5, 6, and 7 are very powerful tools for modifying the original tones and auditory perspective of the instruments recorded in Figs. 2 and 3. The general idea is to provide new sounds that will increase the subjective response to the ultimate product—the record—when it is reproduced.

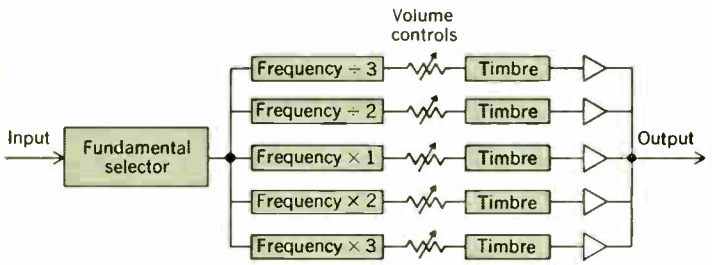
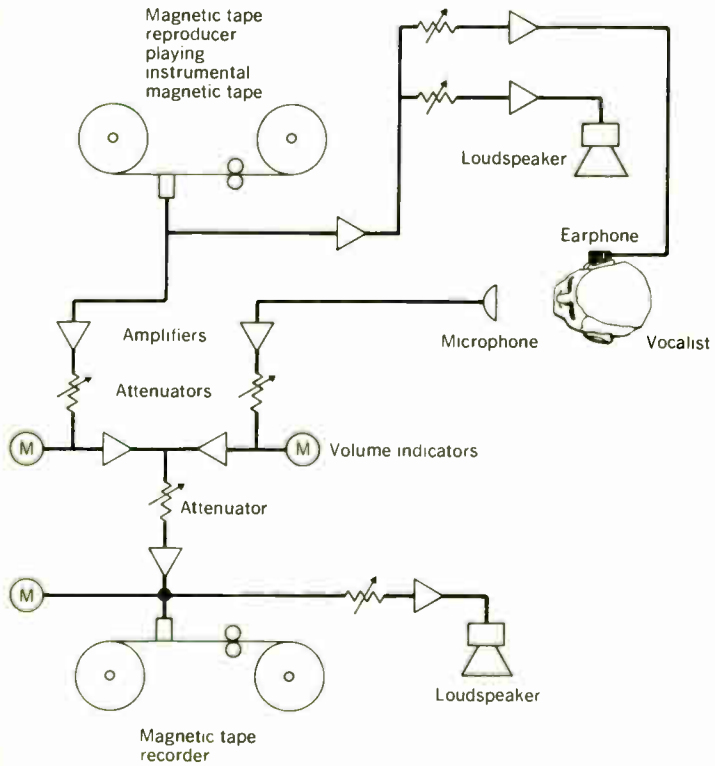


FIGURE 7. A system for selecting the fundamental frequency of an audio signal, multiplying or dividing it by whole numbers, and injecting new timbre to produce several new instruments from the input of one instrument.

FIGURE 8. A method of combining and recording a vocalist with a recorded instrumental number.



In the preceding descriptions, all of the performers and instruments played at the same time. Therefore, the entire unmodified composition was performed in real time. However, in many cases, the vocalist is introduced after the instrumental part has been recorded. This method is illustrated in Fig. 8. The magnetic tape containing the instrumental part is reproduced and fed to the vocalist either by loudspeaker or by earphone. The vocalist then performs along with the instrumental music, and his voice is recorded along with the instrumental part.

The "overdub" process depicted in Fig. 8 is often replaced by the "sel-sync" method, in which the voice is recorded on one track of a multiple track recorder while the other tracks, which have been recorded previously, are reproduced on the loudspeaker or earphones for synchronization. In some recording procedures, the entire instrumental group may not be recorded at the same time. In this case, the "overdub" or the "sel-sync" process is used to combine other instrumental parts at a later time.

The system of Fig. 4 converts eight channels of original sound to two channels of stereophonic sound. In the past few years, practically all records have been produced in the stereophonic form. For the production of monophonic records, however, only one section of the system in Fig. 4 is used. Moreover, by the addition of two more sections of eight channels each (as shown in Fig. 9), the eight channels may be allocated and recorded on four-channel quadriphonic sound. Reverberation envelope, sound in motion, and other spatial effects may be obtained in quadriphonic sound.

All of the conversion of the eight channels to monophonic, stereophonic, and quadriphonic sound, including the modifications, are carried out in a room simulating the acoustics of the average room in a residence. The loudspeaker arrangement for the three systems will be described in the next section.

The preceding exposition, together with Figs. 1 through 9, shows that the combined recording of instruments and performers, along with the modifying processes, in fact constitutes an electronic music synthesis in which the original sounds supply the basic elements but the electronic modifications supply the artistic embellishment so necessary for the production of a recording that will receive a high order of consumer acceptance and, as a result, become a commercially successful product.

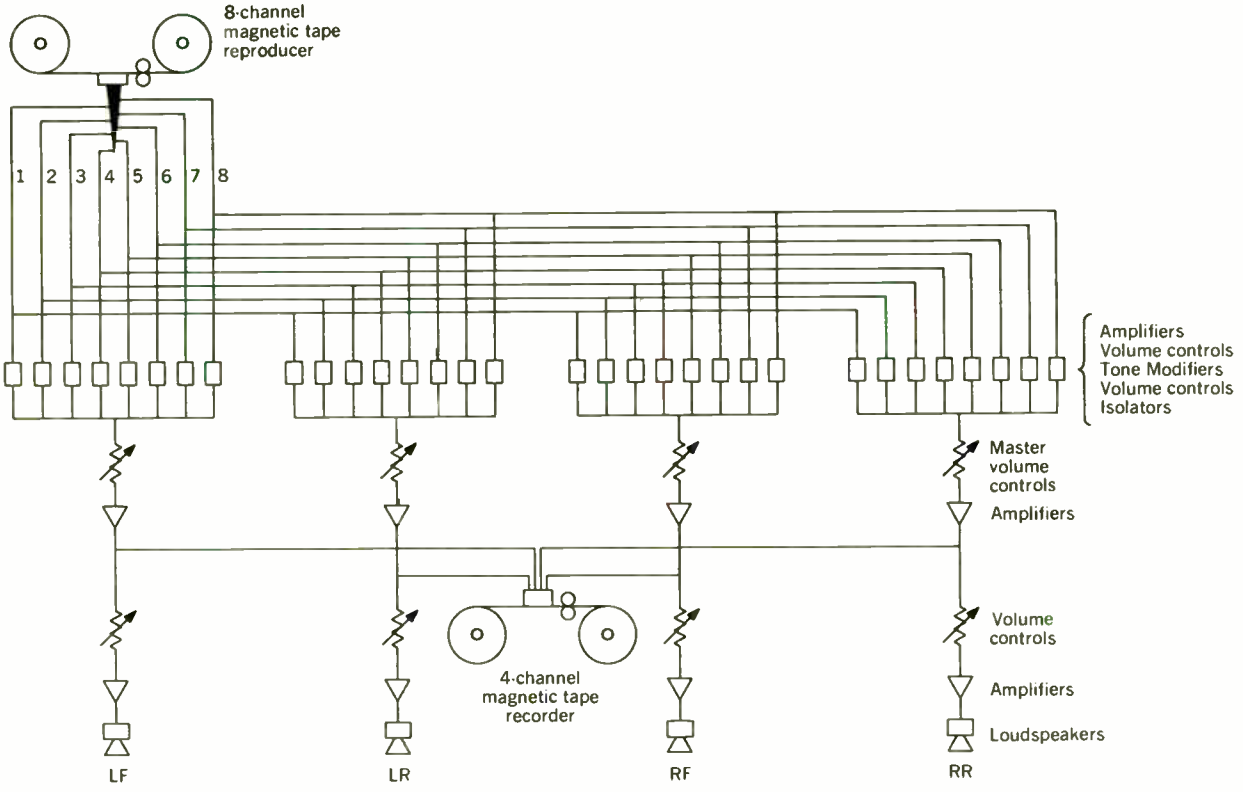


FIGURE 9. The mixing, modifying, and conversion to four-channel quadriphonic sound from the eight sound sources recorded in Figs. 2 and 3.

Monophonic, stereophonic, and quadriphonic sound reproduction systems

Systems for the reproduction of synthesized sound may be classified as monophonic, stereophonic, and quadriphonic. The three systems are described in Fig. 10. The monophonic sound system is of the field type employing a single channel; the stereophonic sound system is of the field type employing two channels; and the quadriphonic sound system is of the field type employing four channels.

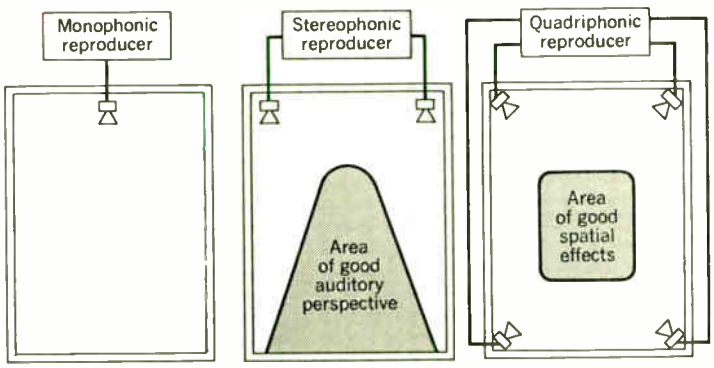
The monophonic sound system can reproduce the entire audio-frequency range with low noise and distortion. In turn, the stereophonic sound system can reproduce the entire audio-frequency range with low noise and distortion and in auditory perspective. Auditory perspective provides the subjective location of the reproduced sound sources. Finally, the quadriphonic* sound system can reproduce the entire audio-frequency range with low noise and distortion, auditory perspective, and a reverberation envelope and spatial effects. Such a system may be employed to provide all manner of acoustic spatial effects such as sound in motion around the room or bouncing back and forth around the room.

Electronic music synthesis has been released as recorded music in monophonic, stereophonic, and quadriphonic form. However, in the past few years practically all records have been produced in stereophonic form. Stereophonic records can also be played on a monophonic reproducer, in which case the two channels are combined into one channel. With the advent of quadriphonic sound, there will invariably be a shift in emphasis from the stereophonic form because of the many additional spatial sound effects.

Synchronized magnetic tape recorders

In motion picture recording and in cases where a single series of tones are produced by a programmed electronic music synthesizer, the recording is made with a synchronized multichannel magnetic tape recorder. In most cases, the magnetic tape is in the form of 35-mm magnetically coated film with sprocket holes as shown in Fig. 11. In this way, the magnetic tape recorder can be synchronized

FIGURE 10. The auditory effects of monophonic, stereophonic, and quadriphonic sound-reproducing systems.



*There has been no standardization of the designation for a four-channel sound reproduction system. However, the terms quadriphonic and quadraphonic appear to be the most common in the literature.

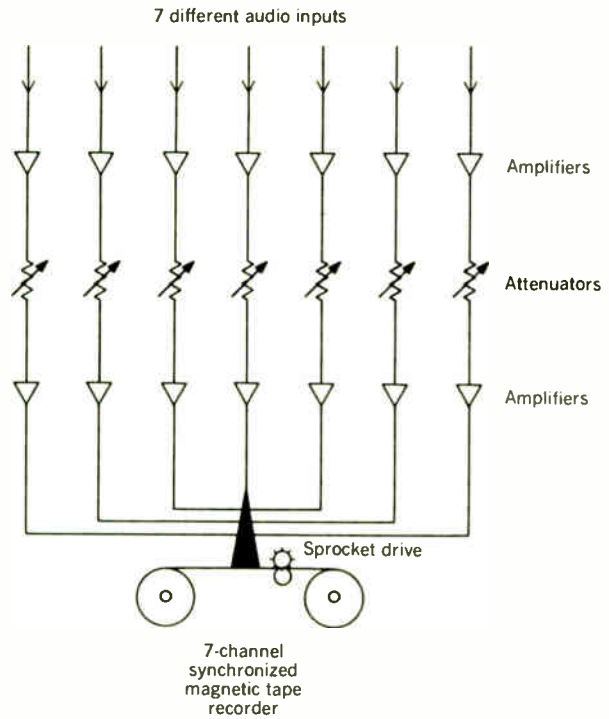
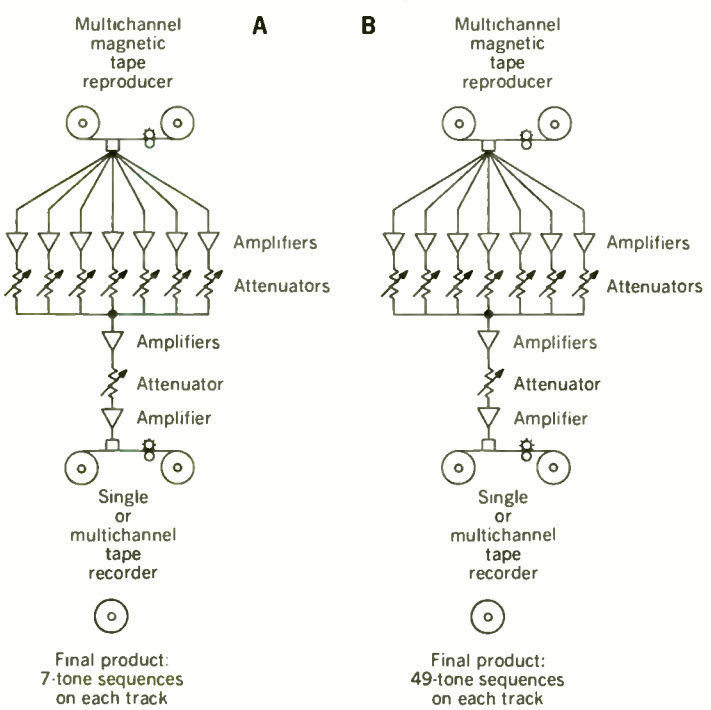


FIGURE 11. Schematic diagram of a synchronized seven-channel magnetic tape recorder for recording seven separate audio inputs or tone sequences on seven tracks of the magnetic tape.

FIGURE 12. A—The seven audio inputs or tone sequences recorded in Fig. 11 are combined into one magnetic track. In turn, if the recorder has a seven-track capacity and all the magnetic tracks are filled, the final product then contains 49 separate audio signals or tone sequences. B—If the seven magnetic tracks of A, each containing seven-tone sequences, are combined to form one magnetic track of the seven-track recorder, then when all the magnetic tracks are filled the final product will contain seven magnetic tracks with 49-tone sequences each, which results in a 343-tone sequence tape. These 343-tone sequences can again be combined into one magnetic track to repeat the process over and over again.



with any picture or program. The synchronized recorder is employed in all types of programmed electronic music synthesis. One design of the synchronized magnetic tape recorder is of the seven-channel type, where seven audio inputs are recorded on seven separate tracks (Fig. 11).

In Fig. 12(A), a magnetic tape with seven separate audio signals recorded in the manner of Fig. 11 is reproduced and the audio signals in the seven channels are mixed and rerecorded on a single track of a seven-track recorder. In turn, if each of the seven tracks of this recorder is itself filled with seven different operations, the total result of the seven tracks if recorded as in Fig. 12(A) represents a 49-tone sequence, as demonstrated in Fig. 12(B). If the recorder in Fig. 12(B) is also capable of seven channels and all the tracks are again filled, the total final result will represent a 343-tone sequence. It is obvious that this procedure can be repeated over and over again, producing tone sequences that are multiples of seven.

Electronic music synthesis by the generation and modification of original sounds[6-9]

Electronic synthesis of music may be created by the generation and modification of all sorts of original sounds, which was the first method used for the synthesis of music by electronic means. A catalog of natural sounds recorded on magnetic tape may also be augmented by the sound created on all kinds of electronic generators. Modifying and shuffling these sounds creates a variety of new sounds that may be classified as discrete units or series of units. These units and combinations of units may be assembled to create musical productions. Hence, there are an infinite number of ways in which electronic synthesis of music may be carried out by the generation and modification of original sounds.

Electronic synthesis of music by these means has been termed *musique concrete* in France, *electronic music* in Germany, Italy, Holland, and Japan, and *tape music* in the United States. The differences in various schools exist in the specific types of approach employed to carry out the process. For example, the source and nature of the original sounds and modification processes are areas in which the various schools differ. It is the purpose of this section to describe the basic process of the electronic synthesis of original sounds.

A generalized diagram describing the main elements involved in mixing and modifying original sources of sound is shown in Fig. 13. The original sources may be recordings on magnetic tape, oscillators, and noise generators. Recordings on magnetic tape may be the sounds of voice, musical instruments, nature, etc. In some cases, the output of more than one magnetic tape reproducer may be mixed (see Fig. 13). The various waveshapes of oscillators may also be employed. For example, the output of an oscillator with a sawtooth waveshape contains the fundamental and all the harmonics; the output of an oscillator with a triangular or rectangular waveshape contains the fundamental and all the odd harmonics. The noise generator is designed to produce white noise and other signals of a random nature.

For our purposes, a mixing system is used to combine the output of the sound sources in the desired arrangement as determined by the musician. The output of the mixing system is then fed to the input of the modifiers. The modifiers include elements for introducing frequency

shift, vibrato, tremolo, portamento, timbre change, growth, steady state, decay, and reverberation.

The frequency shifter is capable of changing the frequency by a fixed amount as in the case of a single-sideband system. If this type of frequency shifting is applied to a signal containing the fundamental and all the harmonics, the overtones will no longer be harmonics. The frequency shifter may change all the frequencies of every component by a multiplicative or divisive factor, with multiple relations of frequency overtones maintained.

The vibrato modulator provides a frequency modulation of the input frequency. The modulation frequency is usually somewhere below 10 Hz.

The tremolo modulator provides an amplitude modulation of the input frequency. The modulation frequency is usually below the audio-frequency range.

The portamento glider supplies a continuous frequency glide in changing from a tone of one frequency to a tone of another frequency.

The timbre modifier consists of frequency-selective networks that accentuate or attenuate the various components of the overtone structure of a complex tone input.

The growth controller determines the growth or attack characteristic of the tone.

The steady-state controller determines the time length of the steady state of the tone after the growth has been established.

The decay controller determines the decay characteristic of the tone. The decay follows the steady-state duration of the tone. The growth, steady-state, and decay controllers are interconnected because the three characteristics follow a direct sequence of events. For example, the steady-state time may be zero, which means that decay of the tone follows immediately after growth.

Reverberation is an artistic (and acoustic) embellishment that plays an important part in blending a series of tones. Subjective considerations have established that for each type of music there is an optimum value of reverberation. A reverberator is included in the system of Fig. 13 to supply artificial reverberation to the sound.

The final product of electronic music synthesis is recorded on magnetic tape or disk for rendition of the music in an auditorium or home.

A rather highly sophisticated system for modifying a tone has just been described for the purpose of illustrating the process. As a matter of fact, innumerable changes can be made in the components and arrangement of the system depicted in Fig. 13 for the electronic synthesis of music by the generation and modification of original sounds, depending upon the composer's requirements.

Formal electronic music synthesizers[2]

In the preceding sections, the electronic synthesis of music has been carried out by modification of conven-

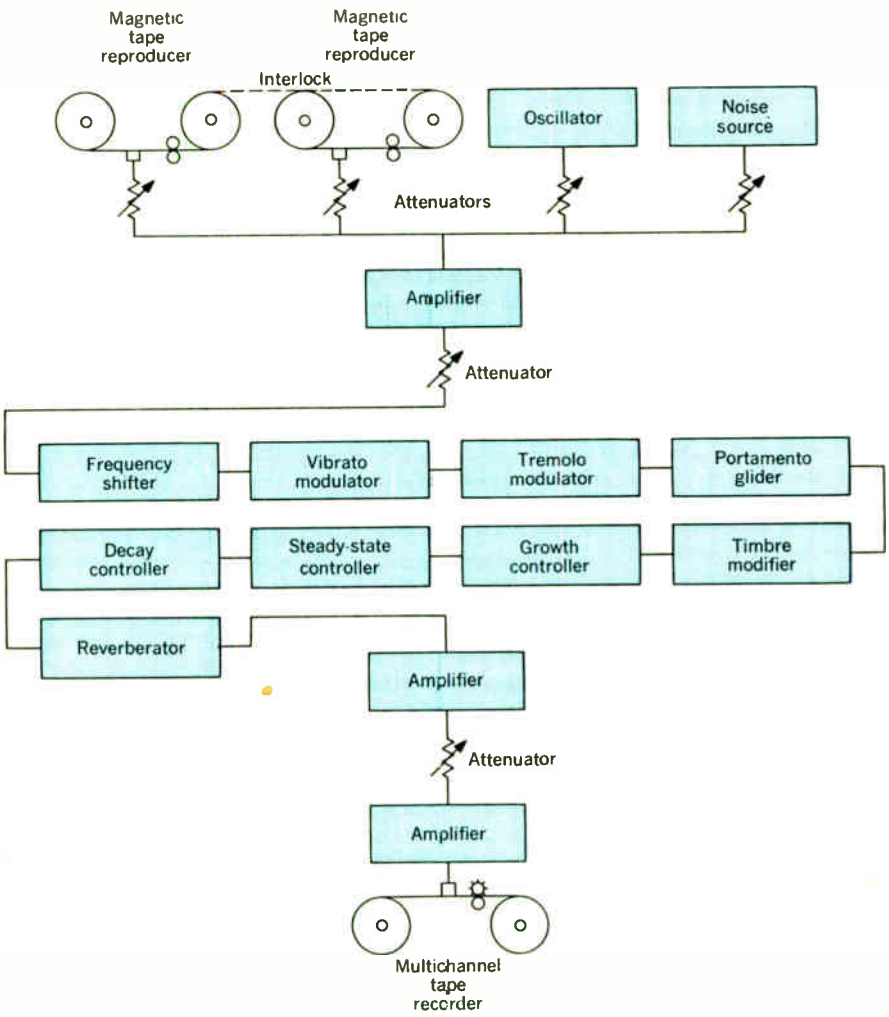


FIGURE 13. Schematic diagram of the apparatus for the production of electronic music by mixing and modifying original sources of sound.

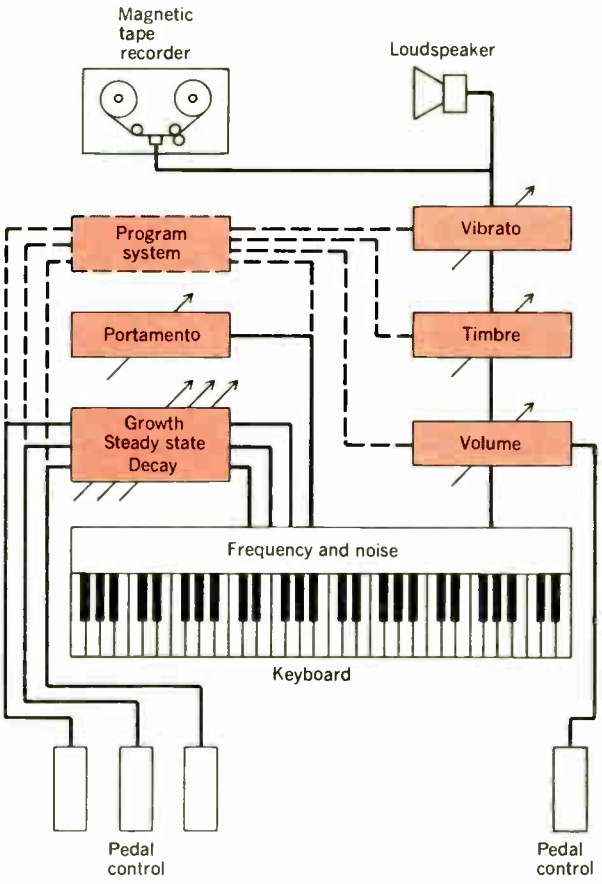
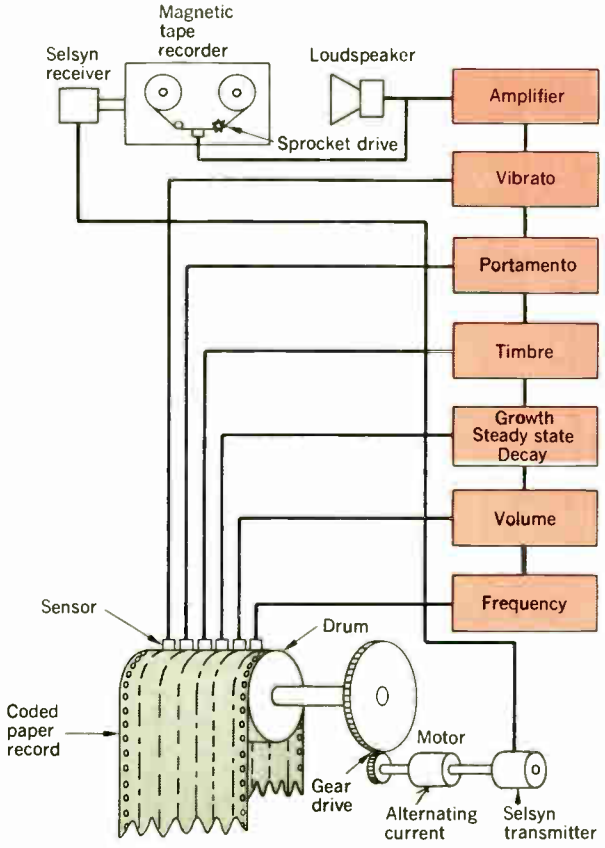


FIGURE 14. The elements of a formal manual electronic music synthesizer.

FIGURE 15. The elements of a formal programmed electronic music synthesizer.



tional and original sounds. The formal electronic music synthesizer, however, as exemplified by the manual, programmed, and computer operation, provides the most sophisticated means for producing recorded music. In such synthesizers, a single series of tones is generated at one time.

Hence, as a result, all the characteristics of a tone (as depicted in Fig. 1) can be applied to each series of tones. This, of course, is impossible in the case of conventional musical instruments or, for that matter, of all electronic musical instruments because, in general, more than one tone is generated at a time. As a matter of fact, the entire composition is produced in one run. Thus it will be seen that synthesizers provide the most powerful system available for the generation of musical sound and the resultant production of recorded music. The sections that follow will provide descriptions of the principal formal electronic music synthesizers.

Manual electronic music synthesizers[10,11]

As we know, the properties of a tone are frequency, intensity, growth, steady state, decay, portamento, timbre, vibrato, and deviation. When these properties of a tone are completely specified, the tone can be completely described. The manually operated electronic music synthesizer, outlined in Fig. 14, is based upon the generation of the these properties to produce the musical tone. The question may be asked as to the difference between a highly sophisticated electric organ and the manual electronic musical synthesizer of Fig. 14. In the synthesizer of Fig. 14, only one series of tones is produced at a time. However, in Fig. 14, each series of tones is not only recorded as in Fig. 11, but the various different series of tones are combined as described by Figs. 11 and 12.

Hence, the synthesizer in Fig. 14 can produce a tremendous range of tones with a maximum variation in each series that cannot be accomplished by the electric organ. For example, the performer is able to change the volume, timbre, and some other characteristics with one hand while he plays on the keyboard with the other hand. The feet are also used to control some of the functions. Therefore, great physical dexterity and considerable musical talent are required to produce acceptable music by means of the manual electronic music synthesizer; the main reason being that each series of tones must possess distinct musical qualities in order to produce musical effects that are beyond the conventional. In addition, there must be a perfect synchronism of the different series of tones that constitute the complete production. Such a synchronizing task is also carried out subjectively, that is, by "ear." It is interesting to note that synthetic reverberation may be introduced either at each note or over the entire composition.

Human engineering has been applied to the manual electronic music synthesizer in order to provide the synthesist with all possible assists in producing each series of tones. Preset functions may be employed for different music events, and various means of interconnection may be provided to augment and extend the manual operation. As the ultimate in assist, program systems[12] are being developed to provide some of the characteristics of tone. Such a program system is indicated by the dashed lines in Fig. 14. The musician plays the keyboard of the instrument and the program system provides the desired tone characteristic at the proper time. The pro-

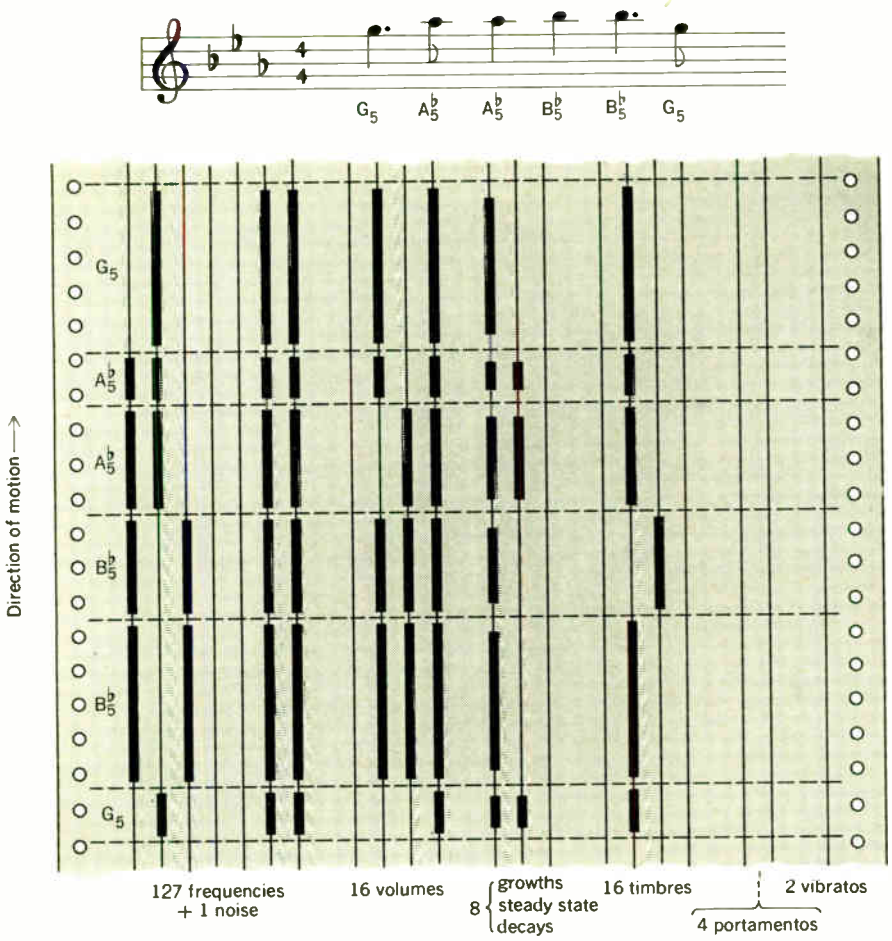


FIGURE 16. Coded paper record for the programmed synthesizer of Fig. 15.

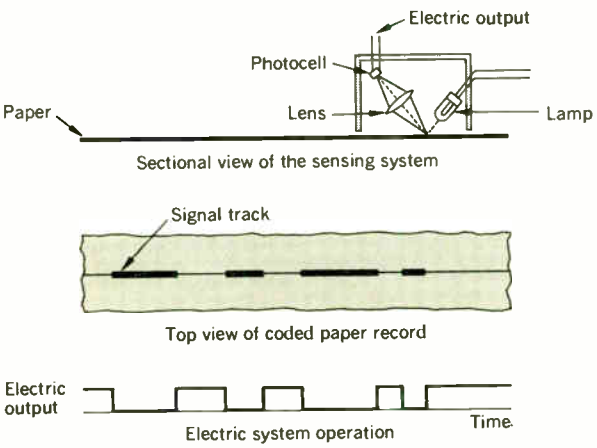
gramming of the system must be carried out by the musician for each series of tones. As probably surmised, technical aspects of synchronizing the programming system and the performing artist become quite complex.

The manual electronic music synthesizer has great appeal for the accomplished performing musician because he can hear the series of tones as he executes them. He can go over and modify the series of tones until he is satisfied with the result, at which time he then carries out the recording process for that series of tones. In terms of variety, manual synthesizers from very simple instruments to exceedingly complex and sophisticated electronic systems have been developed and produced.

Programmed electronic music synthesizers[13–15]

The use of a programmed electronic music synthesizer opens up an entirely new field for the production of recorded music. A programmed synthesizer is detailed in Fig. 15. In order to synthesize any musical tone whatsoever, the synthesizer must possess the following facilities: a means of producing tones with any fundamental frequency within the audio-frequency range; a means of producing tones with any overtone structure; a means of producing a tone with any growth, steady state, or decay characteristic; a means of introducing a vibrato; a means of changing the intensity of the tone; a means of providing a portamento or glide from one frequency to another frequency; and a means of introducing, in the desired instances, various deviations of these characteristics.

FIGURE 17. System for sensing a track of the paper record of Fig. 16. The electrical output represents the signal output of the sensing system for a section of the record.



The program for each series of musical tones is recorded on a coded paper record, which controls all the functions of the electronic music synthesizer. The information is recorded and stored on the paper record in the form of black ink lines* as shown in Fig. 16. When

*The original RCA electronic music synthesizers Mark I and Mark II employ a punched paper record instead of marked paper. The marked paper record is a contribution of D. Friend.

the paper is actually run through the machine, positioned sensors scan the paper record as indicated in Fig. 17. The operation of the sensor is as follows: A lamp illuminates the paper record. When the sensing system passes over a black line, the light received by the photocell (solid state) is reduced and so is the resulting current level. In effect, the sensor is actuated when detecting dark portions of the paper.

It is clear that the paper record of Fig. 16 indicates use of a binary code system. Employing this system, two different functions may be obtained from one track, four from two tracks, eight from three tracks, 16 from four tracks, etc. The gate tree for the binary system is demonstrated in Fig. 18. Here, any one of eight different inputs can be selected by means of three sensors.

Referring to the paper record of Fig. 16, it will be seen that two vibratos are obtained from one track, four portamentos from two tracks, eight growths, steady states, and decays from three tracks, 16 timbres are obtained from four tracks, 16 volumes from four tracks, and 127 frequencies plus one noise from seven tracks. The 88-note player piano requires 88 tracks. With this system, 128 notes can be obtained from seven tracks, clearly demonstrating the saving in paper brought about by the binary system.

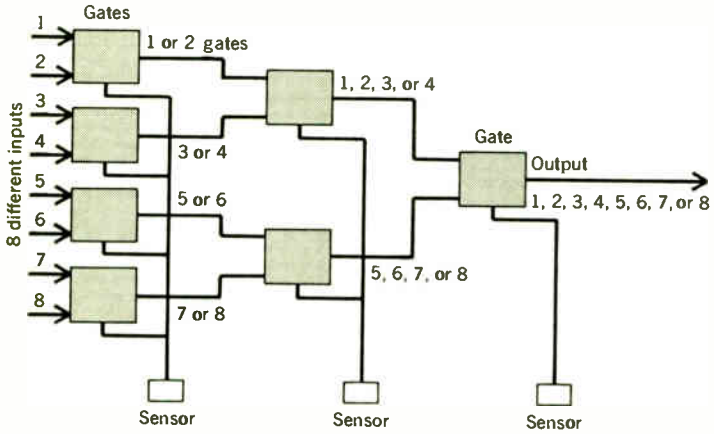
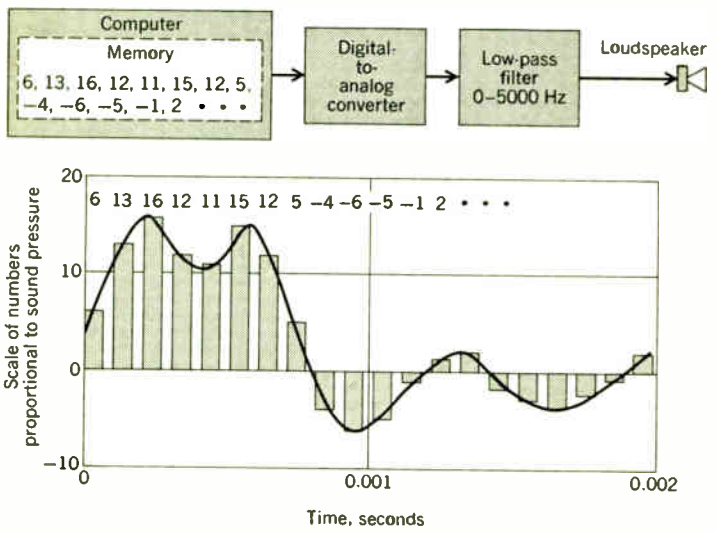


FIGURE 18. A binary sensing system in which any of the eight inputs can be selected by three actuating sensors.

FIGURE 19. Schematic diagram depicting the conversion of a sequence of numbers stored in a digital computer memory into a sound pressure waveform. The sampling rate is 10 000 numbers per second, yielding a bandwidth of 5000 Hz for the sound wave.



The preparation of the record of Fig. 16 is simple. The musician merely marks the paper with a wide pen to obtain the type of sound he desires. This marking of the paper can be done anywhere. After the record is completed, he runs it through the synthesizer and listens to the series of tones by means of a loudspeaker. If he is not satisfied, he can make changes by creating new lines or deleting existing lines with a white marker.

Conversion from the musical score is straightforward, the frequencies being taken directly from the notes on the music score. The synthesist usually knows what he wants in the growth, steady-state, and decay characteristic. The overall interval is determined by the length of the note itself.

In general, the determination of the desired volume is straightforward. This leaves the timbre, which can be determined by trial and error if the synthesist has not carried out work on the electronic music synthesizer before. The portamento is used only in gliding from one note to another in a continuous frequency change. When the vibrato is used, it is a low-frequency modulation of about 7 Hz.

When a series of tones as recorded on the paper record is found to be satisfactory, the result is recorded on one track of a multiple-channel synchronized magnetic tape recorder (see Fig. 15) as used in Figs. 11 and 12. This process is carried out for each series of tones until all of the series of the composition have been completed. Figures 11 and 12 amply describe the method of combining the series of tones. It should be noted that synthetic reverberation is usually introduced throughout the completed composition.

The use of the programmed electronic music synthesizer for the production of musical sounds opens an entirely new field in recorded music. For example, there is the possibility of entirely new tone complexes and combinations that cannot be achieved with conventional instruments. Furthermore, in the case of conventional instruments, the musician is limited to the use of lips, mouth, ten fingers, two hands, and two feet to perform the different functions. In the programmed synthesizer, there are no limitations. If the composer or synthesist has in mind what he wants to achieve, the effects can be obtained by means of the programmed synthesizer regardless of whether he can play a musical instrument or not. In addition, in the case of manual synthesizer, there is the problem of synchronizing the various series of tones. In the programmed synthesizer, however, there is no problem synchronizing these series of tones, as is evident from the direct selsyn drive between the paper record and the magnetic tape recorder (Fig. 15).

Electronic music synthesis by digital computer[16–19]

As already mentioned, music may be synthesized by means of a digital computer. By use of a digital-to-analog converter, the numbers that a digital computer generates can be converted to electrical waves and then sound waves by means of a loudspeaker. The numbers stored in the computer program represent the desired

sound waves to be generated. A schematic diagram of the process of generating sound waves from a digital computer is shown in Fig. 19. The sequence of numbers from the digital computer are converted to pulses of constant width, with the amplitude of the pulses corresponding to the numbers emitted by the digital computer. The pulses are smoothed by a low-pass filter to obtain the input for a loudspeaker (Fig. 19), and a sound wave with frequencies from 0 to N Hz can be generated by $2N$ pulses per second. If the sampling rate is 10 000 per second, the top frequency emitted by the loudspeaker will be 5000 Hz. Assuming each sample is produced from a three-decimal-digit number, the signal-to-noise ratio will be of the order of 60 dB. Hence, within the limits of the frequency range and signal-to-noise ratio, the computer can produce any sound whatsoever provided the appropriate sequence of digital samples can be generated.

The basic procedure of listing 10 000 numbers per second by the composer does provide a high order of generality. However, such a procedure is impossibly tedious and, at the present time, practically out of the question. Furthermore, with the present input means available, such a procedure does not result in a practical solution leading to effective control of the parameters involved in the synthesis of music by means of a computer. Finally, as in the synthesis of music described in preceding sections, the computer synthesis of music must also be based upon the fundamental properties of a musical tone as depicted in Fig. 1.

The basic form of the generating program is a scheme for producing a sequence of sounds representing individual "instruments," which are formed by combining a set of basic building blocks termed unit generators. Appropriate combinations of these unit generators can produce sounds of almost any desired complexity. The compiling program is greatly simplified by the use of macro instructions, which specify a sequence of computer instructions by a single statement. In this way, each unit generator can be specified by a single macro statement.

The first step in producing a musical selection is to punch a set of cards that specify the "instruments" of the "orchestra." By way of example, a very simple orchestra containing only one instrument is represented in Fig. 20. The cards that define both the instrument and the score are typified by Fig. 20(E). Cards 1 through 5 (lines 1–5) define the instrument as two interconnected oscillators, each oscillator having two inputs and one output. The left-hand input specifies the amplitude of the output; the right-hand input, the frequency. The waveshape of the oscillation is specified by stored functions $F1$ and $F2$, which in general are not sinusoidal. The inputs $P5$, $P6$, and $P7$ to the instrument are set by the program to specific values at the beginning of each note. $P5$ determines the amplitude of the note and $P7$ the frequency. $P6$ is set to 1/(note duration); thus oscillator $F1$ goes through exactly one cycle during each note. In this way, $F1$ becomes an attack and decay function that multiplies the waveshape produced by oscillator $F2$, which is a slightly modified square wave whose shape determines the harmonic content of the output waveform.

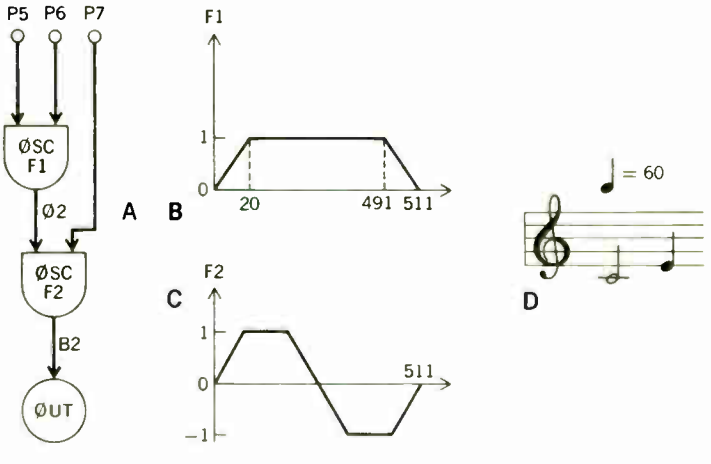
After the instrument has been defined, the next two score cards (6 and 7) cause the functions $F1$ and $F2$ to be calculated and stored in the computer memory. The actual notes are specified by cards 8 and 9. The six numbers on each note card determine the starting time of the note, the instrument on which to play the note, the duration, the amplitude, 1/(duration), and the frequency, in that order. The particular significance of these numbers depends on the particular instrument on which the note is to be played and the composer has a great deal of flexibility in setting up his orchestra. Instruments of greater versatility and complexity may be developed by the addition of vibrato generation, a portamento glider, and other components.

To "play" the music, the computer reads a line from the score, inserting into an "instrument" at the proper time the parameters chosen by the "composer," thus activating the instrument with the generated numbers equivalent to the duration of the note. The sequence of musical events is automatically taken care of by the program and need not concern the composer.

As in the case of the programmed electronic music synthesizer described in the preceding section, the process is simplified if the composer listens to one series of notes at a time, makes suitable modifications, and then groups the combination of tones after each one has been approved. Minor modifications can then be made in the group if such a procedure appears to be desirable.

The description in this section has given one process for the electronic synthesis of music employing a digital computer. As programming for computers is simplified,[20] the general procedure will be simplified. However, the composer operating the synthesizer, regardless of the method employed for the electronic synthesis of music, is the final judge of the rendition. The analog and digital synthesizers can produce any tone whatsoever regardless

FIGURE 20. Instrument with attack and decay. A—Flow diagram. B—Envelope function. C—Waveform function. D—Conventional music score. E—Computer score. (After Mathews)



E

```
 1  INS 0 1 ;
 2  ØSC P5 P6 B2 F1 P30 ;
 3  ØSC B2 P7 B2 F2 P29 ;
 4  ØUT B2 B1 ;
 5  END ;
 6  GEN 0 1 1 0 0 .99 20 .99 491 0 511 ;
 7  GEN 0 1 2 0 0 .99 50 .99 205 −.99 306 −.99 461 0 511 ;
 8  NØT 0 1 2 1000 .0128 6.70 ;
 9  NØT 2 1 1 1000 .0256 8.44 ;
10  TER 3 ;
```

pretty good whereas "probability of failure" has a bad connotation. I have termed this human equivocation the "think-positive syndrome." Some of its aspects have been treated in a recent article.[2]

These two examples may seem trivial at best, childish at worst, but they are neither. They are symptoms of a human problem that seriously interferes with some aspects of good management, and particularly with risk assessment.

Assumptions

The assessment of risk is an essential part of policy development and decision making. All policies or decisions are based on the assessment of the risk involved, whether conscious or subconscious, and whether the assessment is the outcome of a scientific analysis, of intuition, or of both. For this reason the area of the development of information on which to estimate risk is peculiarly sensitive; it is fraught with overdependence of management in its judgment ability, the parochialism of the professions and their disrespect for intuitive values, covert problems of communication, and deeply ingrained preconceptions and prejudices. But today, when all kinds of established views, good and bad, are being questioned, may be an appropriate time to broach the subject.

Since preconceptions are probably a major element in the implementation of risk assessment, I list my own here. I believe that the rest of the article follows reasonably logically from these. However, first it is necessary to have a common understanding of the relation of risk to uncertainty, particularly where the purpose of its assessment is to reduce the risk: *a risk exists because one is uncertain about one's knowledge of some aspect of the problem under consideration.* If one wants to reduce the risk or become more certain of what it consists, it is necessary to recognize and study the uncertainties.

Although both are uniquely related, the term "uncertainties" will be used more frequently than "risk." It is preferred because risk beings to mind a number whereas uncertainty produces a mental image of several individual things and leads one to want to know what they are.

The following list of my preconceptions is focused on conditions that appeared to be unfavorable to the uncovering or communication of uncertainties. They were developed from an inevitably narrow base of personal experiences.

1. *Human judgment.* Except in the simplest cases, such as some games,* risk assessment cannot avoid human judgment.

2. *Models and uncertainties.* A simulation model of a complex situation can be an extremely important analytical tool, but it must not be expected to provide a simple expression for risk. Models are of limited use to a decision maker and can lead him astray unless they clearly express the associated uncertainties that could affect decisions based on them—uncertainties such as boundaries of applicability, assumptions, approximations, and subjective judgments.

3. *Planned reduction of uncertainties.* The relative merits of alternative solutions to a problem cannot rationally be assessed without uncertainties, including the costs and delays needed to reduce them to acceptable levels.

4. *Communication and feedback.* There is no effective communication without some sort of feedback. (The principle applies across interfaces and up and down the organization line.)

5. *Motivation.* There is no consistent and effective activity without motivation.

6. *Environment for motivation.* A sense of achievement combined with understanding and sympathy with ultimate purpose is the most important element of true motivation; it is often even more important than remuneration.

7. *Policies and decisions.* Decisions at all levels lose consistent direction unless they directly or indirectly relate to top-level policies, written or implied, and are supported by top-level action and example sufficient to develop respect and loyalty.

8. *Attention to risk information.* Information on risk is as important to decisions as any other type of project information.

9. *Risk-increasing defects.* Operations that increase risk include:

- Overemphasis on schedule sacrifices risk for efficiency of operation. Underemphasis leads to inefficiency and makes coordinating activities difficult. Proper balance between the two is important.
- Many important decisions, often on what to do about uncertainties, are made at too low a level and seldom surface until too late, or do not surface at all.
- Public recognition or reprimand on routine operations often leads to organized concealment of deficiencies (for example, zero-defect propaganda).
- Management at all levels does not treat uncertainty information with the same diligence and control as it does costs and schedules, although its decisions depend on the assessment of risk. Discussions about the extent to which risk is reduced as the project progresses seldom arise. Little systematic research has been carried out in the difficult area of how to present simply the reduction of uncertainties.
- Managers often seek to control and censor the communication of uncertainties.

Uncovering, communicating, and assessing uncertainties in a routine manner would automatically reduce the risk engendered by these operations.

10. *Systematic approach.* Indeed there is nothing new in the field except a *systematic*, environmental, conscious approach. Presenting and discussing the problem and possible improvements in the management of developing information on uncertainties often produces the reply, "That is just what we are doing" or "We are doing something close to that." But on investigation it is seen that what is being done either is done once in a while as circumstances permit or as someone happens to think about it or actually has little relation to what was suggested be done, or that nothing is being done although some have thought that something should be. There is little indication that thought has been directed toward the environment that would be necessary to nurture the result sought.

This list does not imply that current management practices in the United States are generally bad, for clearly

* Few games are free of human judgment, except for those such as roulette in which the "opponent" operates in accordance with strict quantitative rules. In most games, although the risk appears to be mathematically calculable, judgment is needed in predicting the opponent's reaction (game theory) and is generally based on past experience.

U.S. management has a long history of great achievements. What is indicated is that in the specific and old area of risk assessment, which has suddenly blossomed into an area of major importance, it is peculiarly weak and that many common aspects of management need correction for this purpose. The list also implies that many *positive* things can be done to improve them.

Information and decision

A decision maker has to allow not only for his own uncertainties but also for those of the managers on whom he depends for information and who often will not voluntarily divulge them. The literature is replete with decision makers' negative attitudes on uncertainties, but generally it limits itself to introspective techniques to help them recognize their personal limitations. One finds from Kepner and Tregoe[3] such statements as: "There is a common tendency to overlook the critical consequences of action. Such consequences may be missed because they are too unpalatable to face. . .," ". . .major rewards so often go to those who show the best records of solving current problems in management, and there is rarely a direct reward for those whose foresight keeps problems from occurring. . .," and ". . . [it] is the common conviction of managers and others that any plan of theirs is eminently workable or they would not have suggested it in the first place." Emery[4] refers to the ". . .tremendous vulnerability to emotional bias. Time pressures, old aches from past problems, disparate wishes for improvement and other strong feelings can go to work on vague and questionable data we have about the future that makes systematic evaluation extremely difficult." These psychological impediments occur at all levels, from the man at the bench to the President of the United States.

Similar psychological forces affect the selection and reliability of the information that flows up the line to the decision maker. Top management can do much by providing an environment that relieves its managers from having to struggle with their psychologies by *motivating them to want to uncover and understand uncertainties.*

Temptations of the think-positive syndrome

The sources of the impediments to the uncovering of uncertainties can be lumped under the think-positive syndrome, which is reflected in a common tendency of managers to let "sleeping uncertainties lie." Is it a fault for a manager to react in such a way? Certainly it is, if it affects the quality of his product. But the reaction is such a natural one that it is broadly covered by the unassailable explanation called "human nature." In that guise it tends to be respected and seldom is anything done to eradicate it or reduce its damaging effects.

Let us look at two types of situations in which a manager can find himself. In the first one he is developing a product in a competitive climate. If he expects to be a party in adversary confrontation, whether formal or informal, he knows he will gain nothing by expounding on the uncertainties of his product. Only if he feels that in the confrontation his uncertainties are likely to be uncovered and that he will do better by explaining them himself than by having someone else do so, will he have a personal motivation to study and present them. Even then, the more subtle ones may pass unnoticed so that it does not seem desirable to search more than superficially for these. The brighter the manager, the better he understands these problems—and the greater his repugnance to uncertainties. Can he be blamed if he is not wedded to the concept of ferreting out uncertainties and is less than completely frank in presenting them?

In the second situation the manager is part of a line operation comprising several levels. If he learns about an uncertainty, what alternatives does he have? He can spend time and money assessing and if necessary reducing it, or he can leave it to the next higher level for decision, or he can hope that it will never become noticeable. Since the first two alternatives are not pleasant, and could be very unpleasant, he will tend to be biased in favor of the third, thus taking more risk than is desirable. The probability of personal damage for taking that risk is mitigated, however, by the likely consequences. At one extreme nothing happens: the uncertainty is not noticed and produces no ill effects, or he has been promoted and the ill effects appear under another manager. At the other extreme, the ill effects have accumulated and a crisis has developed. However, rather than being a mark of deficient management, it can provide an opportunity for the manager to show what he can do. The crisis is recognized up and down the line. If he can control it and bring the operation back in a manner satisfactory to the customer, even at a high cost, he is likely to have gained in prestige. Many a middle-level manager has broken through the barrier to higher management, to a place in the "front office," by proven fire-fighting ability. There is little doubt that U.S. managers on the whole are outstandingly competent as fire fighters, but much less so as fire preventers. The conclusion is that current management or organizations do not provide a favorable environment for uncovering uncertainties and therefore for developing unbiased risk assessment.

Criterions for the environment

What then is the environment that would motivate a manager to seek out, obtain, and transmit information on uncertainties? The basic requirement is simple. It is the same as for all business-type activities: *Consistent and systematic uncovering of uncertainties and deciding on appropriate action must accrue to the manager's personal benefit.*

Such a condition exists when the management of risk is an end in itself. It is then the analyst's job to estimate the risk and it matters not to him what he finds the risk to be. This environment exists in many operations—in the business of insurance, for instance, where there is generally no interest in reducing the risk, the risk information being needed to establish the premium to be charged. It exists also for many socially oriented studies where the decision maker's ultimate responsibility may be to seek ways of reducing the risk, but where he gains no personal benefit in presenting a low-risk picture. He should be aware, however, of his own political bias and of the interest and possible bias of his sources of information.

In the case of a line operation, its top manager seeking to improve the development and reliability of uncertainty information and decisions is faced with three alternatives. He can leave the organization substantially as it is, select lower-level managers who are sensitive to the value of risk assessment, and establish a policy or procedure that will uncover uncertainties; or he can take a leaf from manufacturing and set up a parallel organization focusing its attention on uncertainties and reporting not

to the line but directly to him in a manner similar to that of quality control; or he can set up a parallel activity that duplicates much of the work of the line operation as a check on its accuracy.

Let us look at these alternatives.

Environment with unchanged line of operation

First, take the psychologically simpler approach of leaving the organization substantially unchanged. The lower- and middle-management levels are inhibited by the fact that the manager at the next level up (and sometimes higher levels join in) seems unhappy when uncertainties are presented to him; in fact, he generally shows his dislike and concern regarding them. At the end of the analytical line is the decision maker. Basically he has much the same problems as other managers down the line, but often there will be no one in direct authority to feel unhappy about his uncertainties or how he handles them. He is probably constrained, however, by a complex set of political reactions. Sometimes these forces are weak; the feedback delay may be so long that his personal satisfaction in feeling that he is doing a good job in every possible way can overcome the usual destructive effects of the syndrome. He may feel free to dig out uncertainties from those who present alternative programs to him, and give them a sense of achievement for presenting these programs effectively and honestly. If he succeeds he can compare the relative merits of the alternatives, having before him the expected performance, cost, and associated uncertainties, including the cost and delay of reducing them.

Evidently such an environment will stimulate the presentation of uncertainties. Theoretically this can exist at all levels, provided that the manager on the level above makes it abundantly clear that he appreciates it, and that a systematic iterative process is developed by which uncertainties are presented by the lower level for decision by the higher until the uncertainties have been reduced to an acceptable level.

In such an environment the individual's thinking is shifted from fear of the next level up to the satisfaction of achievement. One mandatory requirement is that it start at the top. It will need in addition a period of education, developing understanding of the principle, and a policy of selecting managers whose natural bent fits the motivation for which the environment is being developed. It is believed that generally the more intelligent and better balanced individuals will fit well in comparison with those who are over- or underaggressive.

In view of the present strength of the think-positive syndrome, the uncertainty is not whether the principle can be applied at the top but to what degree it will flow down the line. The question leads us to consider the alternative approach of setting up a parallel organization.

Environment with parallel organization

The obvious method of developing information on uncertainties free from the bias of the undesirable effects of the syndrome is to set up a department for which the most positive action its staff can think about is the uncovering of uncertainties. All that is needed is to set up an uncertainty control manager, reporting directly to top management, with a staff whose duty it is to uncover uncertainties of all kind, assess them, and suggest how they can be reduced and at what cost. Such departments currently exist—for instance, in quality control, reliability, and testing—but these are often weak. They do not all report to top management. Some, instead of controlling, report to the operation that needs their control; their operating personnel frequently have a low status among their professional peers. These unfavorable conditions probably could be remedied by combining all uncertainty-uncovering and analytical functions into one unit, and giving it an important place in the attention of top management and in the official review process. Its influence at reviews is important because it is there that it is decided whether the work has progressed sufficiently and the uncertainties adequately reduced to permit moving to the next stage.

An uncertainty about the effectiveness of such a parallel operation is whether the two parallel lines can work well together at all levels. The two exist to support each other with regard to different aspects of the project but neither one has authority over the other except in the areas of his own specific responsibilities. After some experience the project line may recognize and welcome the value of having uncertainties objectively pointed out. It would reduce the responsibility for mistakes and should end up with a better product. There are many examples of effective mutual support of parallel organizations, so that there is reason to be optimistic about the outcome of this type of uncertainty management.

The most important uncertainty associated with this technique is its cost. The initial reaction to such a process is usually: "We cannot afford it." But is this true? It is difficult to gauge the value of increased understanding of uncertainties and the resulting reduction in the residual uncertainties left at each stage and at the end of a project. Someone has to balance the cost against the expected reduction in risk. It may never be possible to find out whether the introduction of a department of uncertainty control has been cost-effective. Even the cost may be difficult to estimate without practical experience. The cost of the uncertainty line should be a fraction of that of the regular operating line, if mutual support between the lines is effective.

An answer as to whether this approach is worthwhile may come out of experience; after a few applications one may develop a "feel" for its effectiveness. We do have the experience of manufacturing, which may not be identical but certainly is similar to more esoteric operations—and there is no question today but that quality control has fully justified its existence in reducing the risk of defective output early in the manufacturing process. In this case, however, the cost effectiveness is easily measured.

Checks and balances with competing groups

The third alternative is to set up a competing group. When the two groups have completed their work, they compare each other's findings—including uncertainties, establish where differences exist, and try to reduce such differences by analysis and test. The result, including areas of agreement and residual differences, is then presented to the decision maker. With good professional groups the residual differences will generally be small and the decision maker can expect to be faced with a relatively few, well-defined areas requiring decision.

This approach entails two principal uncertainties. The first lies in the inherent problem of adversary confrontation: each group may tend to conceal uncertainties of which it thinks the other may not be aware. With com-

petent groups, however, uncertainties of any importance should be difficult to conceal.

The second, even more important, uncertainty is that the decision maker will tend to accept regularly the findings of one particular group over the other. The "defeated" group or groups gradually will become demoralized and will tend to build an antagonistic attitude toward the decision maker. The "winning" group may become arrogant and develop its own set of prejudices. Antagonism so developed can destroy the intended operation. Such a destructive condition is very likely to take place because one group probably has been selected either because of its exceptional competence or because of its freedom from special interest and bias. This problem will be particularly difficult to overcome when the group that usually loses has divided loyalties—partly to the decision maker and partly elsewhere, perhaps to the organization to which it reports directly. Under such conditions this approach may tend to degrade the relationship between that organization and the decision maker, unless special precautions are taken. An example of the administrative problem involved is to be found in the demise of former Defense Secretary McNamara's "Whiz Kids."

A reason for resistance to this technique is the expense; it may cost twice as much as a single analysis, although in many cases this cost may be reduced because it may be sufficient to duplicate only parts of the analysis.

Conclusion

In the particular area on which this article is focused—the evolution of the information base and its processing for the assessment of risk—the mathematical technology seems far more advanced than its management. In fact, in much of the managerial environment, the cold, self-serving aggressiveness is almost the exact opposite of what is needed to provide the type of environment required for the development of reliable risk information.

However, if one were to transform the present environment into an ideal one for risk assessment, something of value probably would be lost. Possibly two environments can exist side by side,each one covering the area for which it is best suited, but something more important may be involved. The increasing complexity and range of impact of the consequences of policies and decisions is inducing a shift in management priorities. Patterns of management that were satisfactory may need modification. Perhaps, in certain areas at least, new emphases may be evolving under forces arising from the need for better understanding and more accurate evaluation of the consequences of decisions. The increasing need for risk assessment may be one of those forces.

At the present time and for certain missions risk assessment must be bolstered, even at the cost of some other values. It is believed that little or only temporary progress will be made until changes are introduced in the environment that will make the analysis and disclosure of uncertainties accrue to the manager's benefit.

In large multilevel operations three alternatives are open for an environment suitable for risk assessment:

1. Basically unchanged line organization with tight iterative and flow procedures.
2. Installation of an uncertainty control management line parallel to the main analysis line and comprising all uncertainty types of activities.
3. Duplicate organizations with ultimate solution by quasi, well-organized adversary confrontation.

No strong suggestion is made as to which of these is likely to be most satifactory. It will depend on the environment that has existed and can be provided and on the steps that top management is willing to take to provide needed motivation. That motivation will have to compete with existing motivation that may be in conflict with it, including, for instance, analysis of programs with which a department is competing against others for available but limited funds. Generally, therefore, I would prefer alternative 2 because the uncertainties management called for will be judged and will benefit, not by the approval or rejection of a program, but by the correctness and accuracy with which it has uncovered and assessed its uncertainties.

One common requirement mandatory to the success of all three alternatives is total top-management active support.

The discussion has been centered on a particular area. Partly intuitively, partly by considering other areas, it is believed that the problems and principles outlined apply with varying emphases at least to some—and possibly to many—areas of predictive technology.

REFERENCES

1. Schelling, T. C., "PPBS and foreign affairs," paper presented to the Subcommittee on National Security and International Operations of the U.S. Senate Committee on Government Operations, U.S. Government Printing Office, 1968.
2. Wilmotte, R. M., "Engineering truth in competitive environments," *IEEE Spectrum*, vol. 7, pp. 45–49, May 1970.
3. Kepner, C. H., and Tregoe, B. B., *The Rational Manager*. New York: McGraw-Hill, 1965, pp. 208–210.
4. Emery, D. A., *The Compleat Manager*. New York: McGraw-Hill, 1970, p. 82.

Raymond M. Wilmotte (F,L) received the B.A., M.A., and Sc.D. degrees in engineering from Cambridge University in England, and then joined the National Physical Laboratory, the British equivalent of the U.S. National Bureau of Standards, where he was engaged in research on antennas. Arriving in the United States in 1929, he worked on blind landing techniques for the Aircraft Radio Corporation, Boonton, N.J. In 1932, as consultant to the broadcasting industry, he designed, built, and proved out the first directional antenna in the regular broadcast band to protect the service area of a station from the interference of another cochannel station. He has continued consulting throughout his career, first in broadcasting and subsequently with the government through the Wilmotte Laboratory. In 1958 Dr. Wilmotte joined the advanced military systems group of RCA and became program manager responsible for the development and production of the "Relay" spacecraft, the first NASA communications satellite.

During recent years he has been a consultant on special aspects of management to major electronics corporations, nonprofit organizations, and government agencies. He received the Bureau of Ordnance Development Award from the Navy Department for his efforts during World War II, has published over 50 professional papers, and has been granted more than 40 patents.

Linear circuit applications of operational amplifiers

Operational amplifiers are based on the principle of using dc amplifiers employing feedback in order to control response characteristics for the purpose of performing various mathematical operations

Larry L. Schick *Burr-Brown Research Corporation*

Some of the most frequently encountered linear circuit applications of operational amplifiers are discussed in this article. Included are details of differential dc amplifiers, bridge amplifiers, analog integrators, differentiators, line-driving amplifiers, ac coupled feedback amplifiers, current-to-voltage converters, reference-voltage sources, voltage regulators, current amplifiers, and charge amplifiers.

Differential dc amplifiers

The amplifiers to be discussed in this section are most descriptively known as differential dc amplifiers, denoting the fact that they amplify the difference between two signals, and that the inputs are directly coupled. Other common terms used for this basic type of amplifier are: transducer amplifier, bridge amplifier, data amplifier, instrumentation amplifier, difference amplifier, and error amplifier. Such amplifiers are easily realized through the use of one or more operational amplifiers with linear feedback. The idealized characteristics include infinite input impedance, zero output impedance, no dc offsets or drift, zero amplifier noise, a constant gain factor with no gain error, and complete rejection of signals common to both inputs (infinite common-mode rejection). Inputs are typically from transducers, which convert a physical parameter and its variations to electric signals. Examples of such transducers include thermocouples and strain-gage bridges. Several types of such differential dc amplifiers, of varying complexity and performance characteristics, are discussed in the following paragraphs.

Differential dc amplifiers using one operational amplifier.[1-3] The circuit of Fig. 1(A) has the virtue of simplicity, in that it uses only one operational amplifier and four matched resistors. The presence of a common-mode voltage e_{cm} and a differential voltage $(e_1 - e_2)$ is characteristic of most transducers. The common-mode voltage may represent a dc level, as in a bridge, or noise pickup. If an ideal operational amplifier is assumed, the following equations apply:

$$e_3 = (e_{cm} + e_2)\left(\frac{R_4}{R_3 + R_4}\right)$$

$$\frac{e_{cm} + e_1 - e_3}{R_1} = \frac{e_3 - e_o}{R_2}$$

$$e_\epsilon \approx 0$$

where e_ϵ is the amplifier input voltage.

Combining the first two of these equations gives the resulting equation for output voltage:

$$e_o = e_{cm}\left[\frac{R_4R_2 + R_4R_1 - R_2R_3 - R_2R_4}{R_1(R_3 + R_4)}\right] - \frac{R_2}{R_1}e_1 + \left(\frac{R_4}{R_3}\right)\frac{1 + (R_2/R_1)}{1 + (R_4/R_3)}e_2$$

If $R_2/R_1 = R_4/R_3$, the foregoing equation reduces to

$$e_o = \frac{R_2}{R_1}(e_2 - e_1)$$

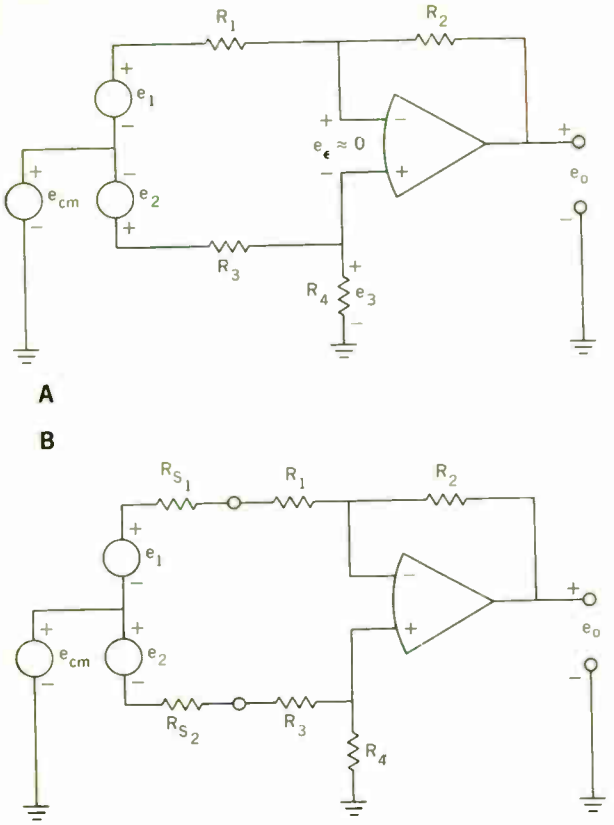
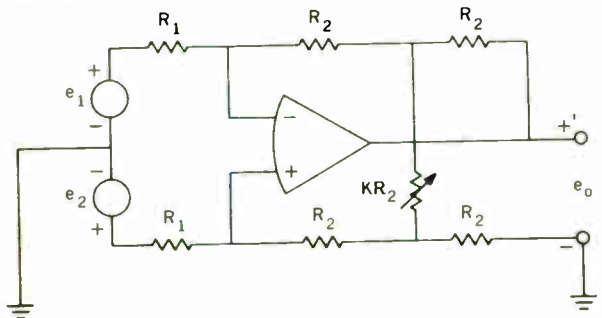


FIGURE 1. Simple differential amplifier. A—Zero source impedance. B—Unbalanced source impedance.

FIGURE 2. Simple, adjustable-gain differential amplifier.



The resistor ratios, R_2/R_1 and R_4/R_3, must be carefully matched in order to insure the rejection of common-mode signals. The value of these resistor ratios sets the gain for differential signals. These equations illustrate the performance of the circuit when dealing with zero source impedances and nonzero common-mode signals. For zero source impedance the gain is determined solely by the feedback resistors and, if these resistors are matched in pairs as indicated, common-mode signals are rejected completely. Actually, of course, the operational amplifier has been assumed ideal in having infinite input impedance, infinite gain, and infinite common-mode rejection. If these factors are given real values and their effects evaluated, it will be found that the finite input impedance of the operational amplifier and its inherent finite common-mode rejection will place limits on the overall common-mode rejection of the closed-loop differential amplifier. The finite open-loop gain will limit the gain accuracy of the overall circuit.

Figure 1(B) illustrates a model for unbalanced source impedances and their interactions with the finite resistances of the amplifier feedback network. An analysis similar to that for the circuit of Fig. 1(A) yields

$$e_o = e_{cm}\left[\frac{R_2(R_{S_1} - R_{S_2})}{(R_1 + R_{S_1})(R_3 + R_{S_2} + R_4)}\right] + \left(\frac{R_2}{R_1 + R_{S_1}}\right)\left[\frac{1 + (R_1 + R_{S_1})/R_2}{1 + (R_3 + R_{S_2})/R_4}\, e_2 - e_1\right]$$

Note that if the source impedances are nonzero but equal, the only effect is a gain error due to the source loading. However, if the source impedances are also unequal, the common-mode rejection is degraded.

Input bias currents (I_{B1}, I_{B2}) and input voltage offset (V_{OS}) of the operation amplifier will cause dc offset errors at the output of the differential amplifier circuit. Bias current (I_{B2}) from the noninverting side of the operational amplifier flows through the parallel combination of R_4 and R_3 to create a dc error voltage at the noninverting input terminal. This dc voltage effectively adds to the offset voltage of the operational amplifier and is amplified by the factor, $(R_2 + R_1)/R_1$. Bias current (I_{B1}) from the inverting input of the operational amplifier flows principally through resistor R_2 and causes an output offset adding to the other two components to give the total dc offset error of

$$E_{OS} = V_{OS}\left(\frac{R_2 + R_1}{R_1}\right) + I_{B1}\left(\frac{R_3R_4}{R_3 + R_1}\right)\left(\frac{R_2 + R_1}{R_1}\right) - I_{B2}R_2$$

Tracking between the two bias currents reduces the bias current induced error term by as much as a factor of 10.

The principal limitations of this circuit are its low input impedance and the difficulty of varying the gain. The input impedance, of course, is determined by the feedback and input resistors. If these resistors are made large in order to increase the input impedance, the dc errors due to bias currents will be proportionately increased, thus placing an upper limit on the feasible values of input impedance. The gain of the differential amplifier can be changed only by varying the ratios of the feedback resistors. Because of the necessity for maintaining the equality of the resistive ratios, it is quite difficult to vary the gain continuously. Gain steps can be achieved if the common-mode rejection is carefully adjusted at each gain setting.

The differential amplifier circuit of Fig. 2 is a similar type of circuit, with the added feature of a gain vernier that allows the gain to be continuously varied without affecting the common-mode rejection of the circuit. The output voltage is

$$e_o = 2\left(1 + \frac{1}{K}\right)\frac{R_2}{R_1}(e_2 - e_1)$$

Note, however, that this circuit requires four matched resistors of value R_2 and two matched resistors of value R_1. The gain is an inverse function of the setting of the vernier potentiometer and, as such, is highly nonlinear. The potentiometer can, however, provide approximate

linearity over limited ranges. The circuit still suffers from the limitations of low input impedance. The dc offset errors are much the same as those for the circuit of Fig. 1.

Differential dc amplifiers using more than one operational amplifier.[1,3] The circuit of Fig. 3 provides another low-impedance alternative to those of Figs. 1 and 2. The two amplifiers required operate in the inverting mode and need not have a noninverting capability. Thus they can be chopper-stabilized amplifiers for low drift or may be field-effect transistor (FET) input types, which may have rather poor linearity when used noninverting. The output voltage is

$$e_o = \frac{R_2}{R_1}(e_2 - e_1)$$

The gain can be easily varied, in steps or continuously, by changing the value of R_2, without affecting the common-mode rejection properties. Good common-mode rejection requires four closely matched resistors of value R_1. Note that the dc offset error is approximately four times that of a single amplifier, if it is assumed that offset errors add, as given by the expression

$$E_{os} = \left(1 + 2\frac{R_2}{R_1}\right)V_{os2} + 2\frac{R_2}{R_1}V_{os1}$$

$$= \left(2 + 4\frac{R_2}{R_1}\right)V_{os} \text{ (worst case)}$$

Since the common-mode rejection of the operational amplifiers is not a factor, the common-mode rejection of the closed-loop amplifier can be trimmed to quite high values by simply allowing a small amount of adjustability of one of the R_1 resistors. The common-mode voltage capability of the circuit is limited only by the output voltage capability of the unity gain inverter. This capability can be increased by making the gain of amplifier 1 less than unity. The gain of amplifier 2 must then be increased accordingly, however, thereby increasing the output offset error.

Another differential dc amplifier circuit using two operational amplifiers is shown in Fig. 4. This circuit provides the high input impedance lacking in the circuits discussed up to now. For this circuit,

$$e_o = \left(1 + \frac{R_4}{R_3}\right)(e_2 - e_1) \qquad \text{if } \frac{R_1}{R_2} = \frac{R_4}{R_3}$$

Again, equality of the two resistor ratios is required if the circuit is to reject common-mode signals. The operational amplifiers, since they operate in the noninverting mode, must have good common-mode properties. The input impedance at each terminal of the differential amplifier is simply the common-mode input impedance of the operational amplifiers. This can be quite large (10 MΩ and up) depending on the type of operational amplifier used. For fixed gains, or gain steps, the circuit is quite useful, but it is not feasible for continuously variable gain. Also, since the input voltage of the upper amplifier must be less than $R_1/(R_1 + R_2)$ times the output saturation voltage, the common-mode voltage range is very limited at low values of overall gain. This is not considered a serious limitation, since such amplifiers are usually used at gains of 10 or greater.

The differential dc amplifier circuit of Fig. 5 overcomes most of the weaknesses of the circuits discussed up to this

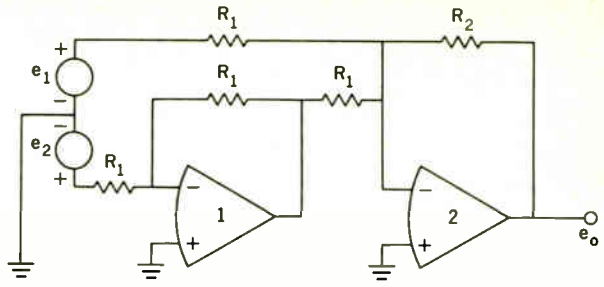


FIGURE 3. Differential dc amplifier using inverting operational amplifiers.

FIGURE 4. High-input-impedance differential amplifier.

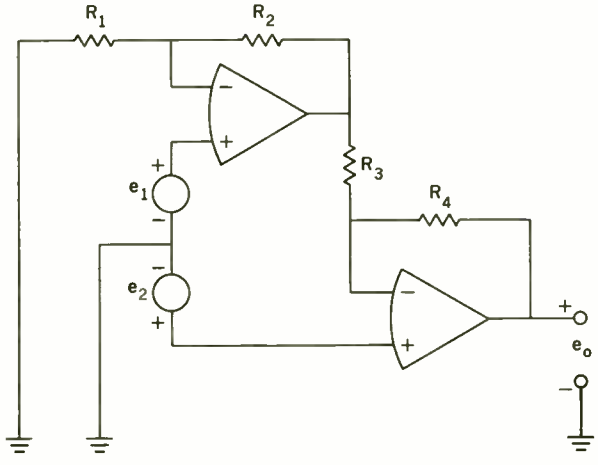
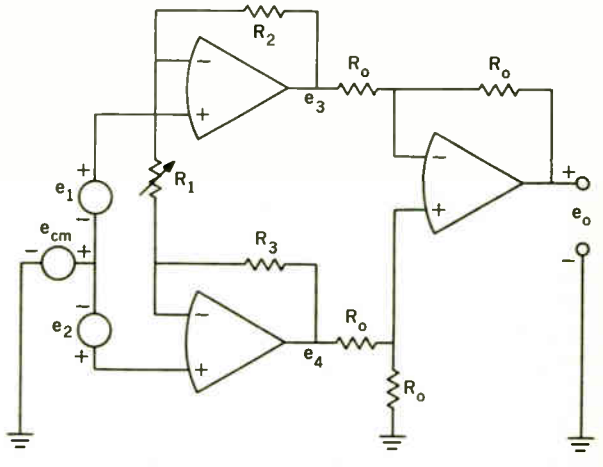


FIGURE 5. High-input-impedance adjustable-gain differential amplifier.



point. Analysis of the circuit yields the following equations:

$$e_3 = \left(1 + \frac{R_2}{R_1}\right)e_1 - \frac{R_2}{R_1}e_2 + e_{cm}$$

$$e_4 = \left(1 + \frac{R_3}{R_1}\right)e_2 - \frac{R_3}{R_1}e_1 + e_{cm}$$

$$e_o = e_4 - e_3$$

If $R_2 = R_3$, then the output voltage is

$$e_o = \left(1 + \frac{2R_2}{R_1}\right)(e_2 - e_1)$$

The two input amplifiers constitute a differential buffer amplifier with a gain of $1 + (2R_2/R_1)$ for differential signals, and unity gain for common-mode signals. The noninverting configuration of these input amplifiers insures high input impedance at both inputs. The gain is easily varied by a single resistor R_1. The effect of mismatch in resistors R_2 and R_3 is simply to create a gain error without affecting the common-mode rejection of the circuit. The resistors R_a of the output amplifier must be accurately matched, or trimmed, to insure the rejection of common-mode signals at this point. This final amplifier acts simply as a differential-input/single-ended-output converter. Feedback impedances in both stages can be relatively low in value to minimize the effects of bias current, since these feedback elements do not affect the input impedance of the differential amplifier. Usually, all of the gain of this differential amplifier is in the input stage, thus insuring that only the offset voltages of these two operational amplifiers are significant in determining the output offset. Since the output voltage offset is proportional to the difference of the voltage offsets of these two amplifiers, it is desirable to use amplifiers whose voltage offsets tend to track with temperature. Such techniques are the basis for some low-drift differential-amplifier modules. The bias currents of these input amplifiers will flow through the impedance of the source, and will thus generate additional offset voltage, which will appear at the output of the differential amplifier amplified by the differential gain factor. The use of amplifiers with FET input stages will greatly reduce this effect.

Bridge amplifiers[1]

Probably the most common use for a differential dc amplifier is in amplifying the output signal from a transducer bridge, such as a strain gage. The most straightforward way of doing this is with one of the high-impedance amplifiers discussed in the previous section. Such a strain-gage bridge, with one active bridge arm, is shown in Fig. 6. The following equations describe its operation:

$$e_2 = V\frac{R}{2R + \Delta R} \qquad e_1 = \frac{V}{2}$$

$$e_2 - e_1 = -\left(\frac{V}{4}\right)\frac{\delta}{1 + (\delta/2)}$$

where $\delta = \Delta R/R$. Then

$$e_o = K(e_2 - e_1) = -\left(\frac{KV}{4}\right)\frac{\delta}{1 + (\delta/2)}$$

or

$$e_o \approx -KV\frac{\delta}{4} \qquad \text{if } \delta \ll 1$$

The output signal is a linear function of the variation of the active element only for small percentage changes in the element. If larger changes are to be measured, the exact equation must be used and a conversion or linearization must be performed at some point in the data-gathering process.

It is sometimes desirable to use an amplifier less complex than the fully developed differential instrumentation amplifier for amplifying the output signal from a bridge.

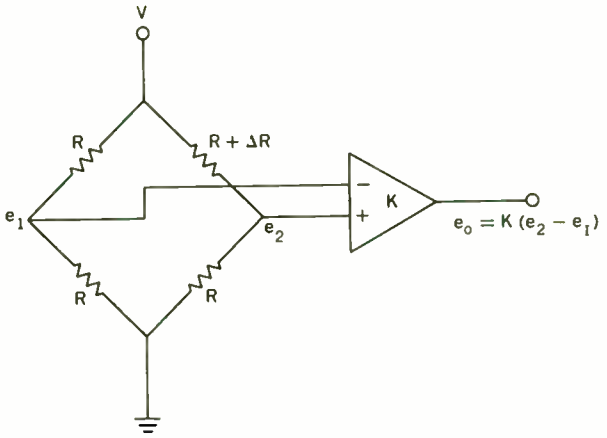
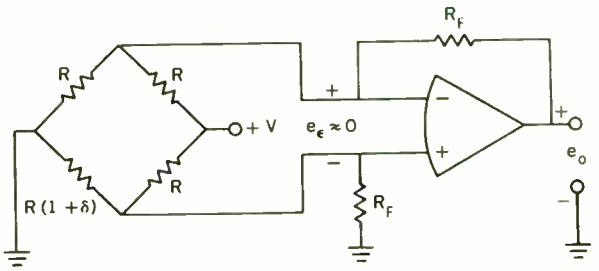


FIGURE 6. Bridge amplifier.

FIGURE 7. Bridge-type current amplifier.



There are several such circuits that use only a single operational amplifier, such as the one shown in Fig. 7. This circuit forces the differential output voltage of the bridge to be zero, since opposite sides are connected directly to the inputs of an operational amplifier with feedback. Thus the amplifier is used to measure the current flowing into the bridge under short-circuit conditions. The resulting output voltage is

$$e_o = \frac{R_F}{R}\left(\frac{\delta}{1 + \delta}\right)\frac{V}{\left(\frac{2 + \delta}{1 + \delta} + \frac{R}{R_F}\right)}$$

If $\delta \ll 1$ and $R_F \gg R$, this equation reduces to the approximate form

$$e_o \approx V\left(\frac{\delta}{2}\right)\frac{R_F}{R}$$

Note that here again the equation for the output voltage of the bridge amplifier is a nonlinear function of the variation of the active bridge element, but for small deviations the nonlinearity is negligible. For the simplified, approximate form of the equation, it has also been assumed that the values of resistance in the bridge are much smaller than the resistors R_F. The bridge resistance appears in the gain equation, thus requiring that the values of the bridge elements be insensitive to temperature in order that the gain of the amplifier be stable with temperature. If the assumption that R_F is much greater than the nominal bridge resistance applies, there is no loading effect.

The dc offset voltage generated at the output of the bridge amplifier as a result of the input offset voltage

and bias currents of the operational amplifier is given by

$$E_{OS} = V_{OS}\left(\frac{2R_F + R}{R}\right) + (I_{B2} - I_{B1})R_F$$

where I_{B1} and I_{B2} are input bias currents. The main advantage of this circuit is its simplicity. It does require an amplifier with reasonably good common-mode rejection.

Where the rejection of common-mode noise signals is not a problem, the half-bridge measuring circuit of Fig. 8 is sometimes used. Here, also, the output of the bridge is connected directly to the input terminal of the operational amplifier as is the feedback through R_F. Since the other input of the operational amplifier is held at ground potential, the output of the half-bridge is held at zero voltage, and the amplifier responds to the short-circuit output current

$$e_o = -iR_F = -V\frac{R_F}{R}\left(\frac{\delta}{1+\delta}\right)$$

If $\delta \ll 1$,

$$e_o \approx -V\left(\frac{R_F}{R}\right)\delta$$

Because the amplifier does operate single-ended, the amplifier used can be chopper-stabilized for lowest possible drift and dc offset errors. Also, the maximum voltage supplied to the bridge, or half-bridge, is not limited by common-mode voltage limitations of the operational amplifier, as it is in those circuits that use the noninverting input of the operational amplifier. Thus, it is possible to increase the sensitivity of the bridge by increasing the supply voltage within the limitations of the bridge elements and the ability of the amplifier to supply the current flowing through the feedback resistor.

The major drawback of the half-bridge circuit is its inability to reject noise pickup, as is normally accomplished by the differential type of bridge amplifier. Consequently, the noise and ripple of the half-bridge supply must be very low and all wiring must be kept short and well-shielded. As in the previous bridge amplifier, the gain is a function of the bridge elements. This can be a serious drawback if the bridge elements are sensitive to environmental factors other than the one that it is desired to measure. The output dc offset voltage of the half-bridge amplifier is given by the expression

$$E_{OS} = V_{OS}\left(1 + \frac{2R_F}{R}\right) - I_{B1}R_F$$

where I_{B1} is the input bias current.

Figure 9 illustrates another bridge amplifier using a single operational amplifier in the inverting mode. Thus it is once again possible to use a single-ended chopper-stabilized amplifier with its attendant low drift. The amplifier output voltage is

$$e_o = V\left(1 + \frac{R_F}{R_1}\right)\frac{\delta}{4[1 + (\delta/2)]}$$

which, for $\delta \ll 1$, reduces to

$$e_o \approx V\left(1 + \frac{R_F}{R_1}\right)\frac{\delta}{4}$$

Another advantage of this circuit, not shared by the preceding two, is that the gain is not dependent upon the

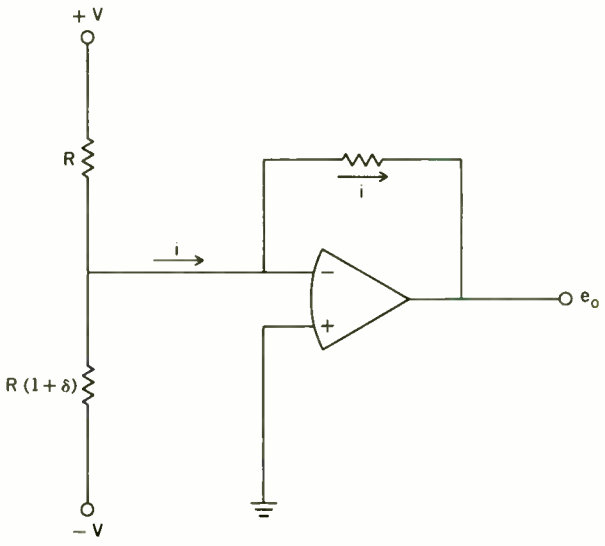
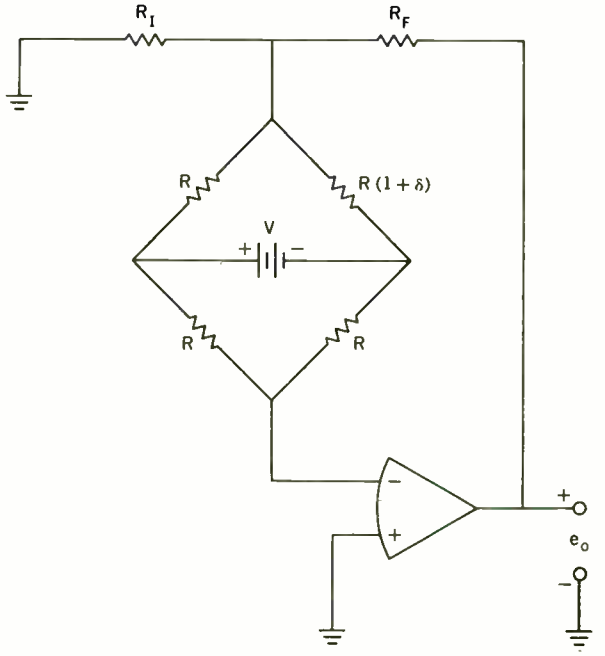


FIGURE 8. Half-bridge current amplifier.

FIGURE 9. Inverting bridge amplifier.



absolute value of the bridge resistors. The output voltage is proportional to the open-circuit voltage of the bridge since the input to the amplifier draws negligible signal current. The inverting input of the operational amplifier is maintained at virtual ground by the high open-loop gain. Since the gain is a function of R_F and R_1, it can be varied easily with either resistor. A small-valued potentiometer can be added in series with either resistor for calibration purposes. This type of bridge amplifier, which can be very accurate, is recommended when it is necessary to detect very small bridge signals. The primary disadvantage is that a floating bridge supply is required. Since it uses a single-ended amplifier it does not have the common-mode-rejection capabilities of the true differential amplifier. However, careful shielding and filtering to remove noise can help to eliminate this prob-

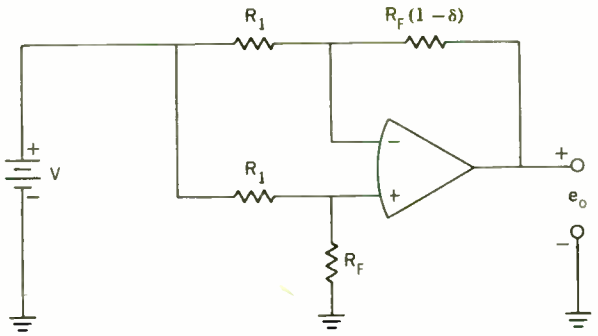


FIGURE 10. Wide-deviation bridge amplifier.

FIGURE 11. Analog integrator.

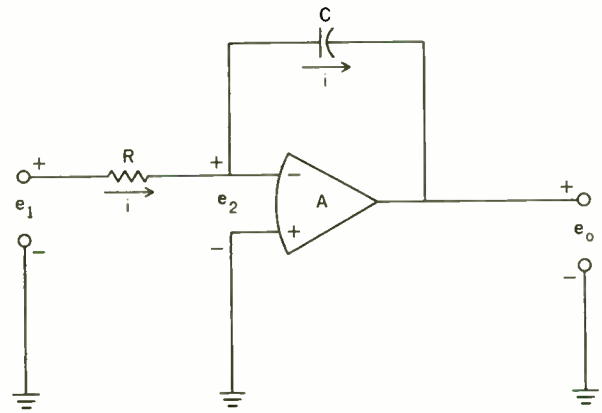
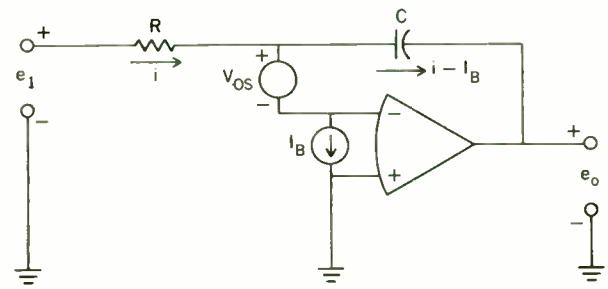


FIGURE 12. Effects of offset voltage and bias current in integrator circuit.



lem. The output offset voltage as a function of input offset voltage and bias currents is similar to that of the inverting amplifier circuit; that is,

$$E_{OS} = \frac{R_F + R_1}{R_1}(V_{OS} - I_{B1}R) - I_{B1}R_f$$

The final bridge-amplifier circuit to be discussed is that given in Fig. 10, in which the output voltage is directly proportional to the transducer deviation even for large fractional changes in the active element; that is,

$$e_o = -V\left(\frac{\delta R_F}{R_1 + R_F}\right)$$

This particular circuit should be used whenever the deviation of the active element is large enough that the linear approximations made in the previous bridge equations are no longer valid. Examples are thermistors, semiconductor strain gages having high gage factors, etc. The bridge elements must be so matched that the two input resistors are equal and the active element is equal to the value of R_F when the bridge is at null. Calibration is difficult since it requires the trimming of two values of resistance to maintain null while varying sensitivity.

Analog integrators[4-6]

The analog integrator is extremely useful in computing, signal-processing, and signal-generating applications. It uses an operational amplifier in the inverting configuration, as shown in Fig. 11. The equations of operation are derived assuming an ideal operational amplifier of gain A. These are

$$\frac{e_1 - e_2}{R} = i$$

$$e_2 - e_o = \frac{1}{C}\int_0^t i\,dt = \frac{1}{RC}\int_0^t (e_1 - e_2)\,dt$$

$$e_2 = -\frac{e_o}{A}$$

If $A \to \infty$, then $e_2 \to 0$, and

$$e_o = -\frac{1}{RC}\int e_1\,dt$$

As in the inverting amplifier, the summing point is held at a virtual ground by the high gain of the amplifier and its feedback network. Since no current flows into the input terminal of the operational amplifier, all of the input current, $i = e_1/R_1$, is forced to flow into the feedback capacitor, causing a charge voltage to appear across this element. Because one end of the capacitor is tied to the virtual ground point, the output voltage of the amplifier equals the capacitor-charging voltage. The overall integrator circuit has the low output impedance normally associated with a feedback amplifier.

The dc offset and bias current of the analog integrator are taken into account in the more realistic model of Fig. 12. Because these dc errors exist, the output of the integrator now consists of two components: the integrated signal term and an error term

$$e_o = -\frac{1}{RC}\int e_1\,dt + \frac{1}{RC}\int V_{OS}\,dt + \frac{1}{C}\int I_B\,dt + V_{OS}$$

The error term itself is made up of a component due to the input offset voltage and another due to the input bias current. The integral of the dc offset voltage results in a ramp voltage, a linearly increasing term whose polarity is determined by the polarity of the input offset voltage. In addition to this ramp-voltage error, the input offset voltage creates an output offset voltage equal to it in value. The bias current flows almost entirely through the feedback capacitor, charging it in ramp fashion, similar to the ramp voltage resulting from the input offset voltage. These two ramp-voltage errors will continue to increase until the amplifier reaches its saturation voltage or some limit set by external circuitry. These error components usually set the upper limit on feasible length of integration time. The error component caused by bias current can be minimized by increasing the capacitance of the feedback element. This can be done only by de-

creasing the value of the input resistor, if a specific value of the RC time constant is to be achieved. A lower limit usually exists on R because of current limitations and loading of the input signal source.

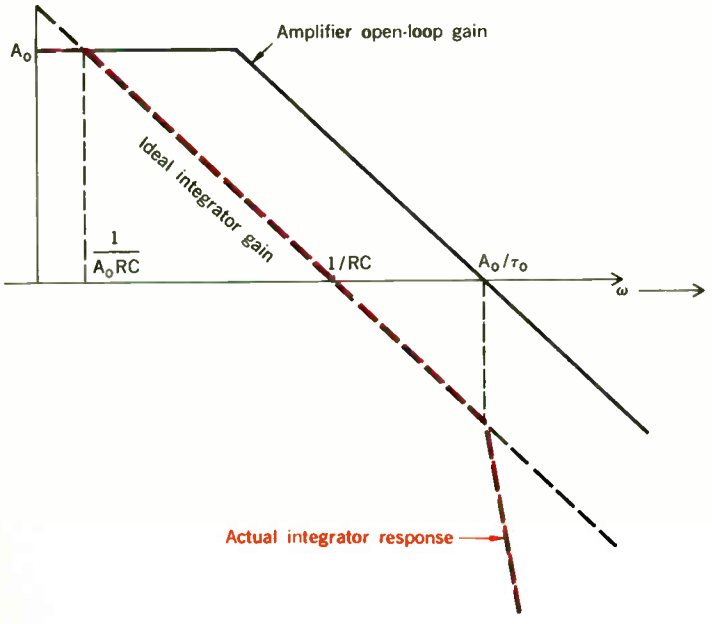
The effects of bias current can be reduced by inserting a resistance R between the noninverting input of the amplifier and ground. This equalizes the resistances at the two inputs and changes the effects of bias current to that of offset (difference) current. Thus, in the equation for output voltage, the bias current I_B should be replaced by the offset current I_{OS} if the compensating resistor is used. The error ramp due to voltage offset is fixed by the chosen value of RC time constant.

To realize the performance possibilities of an operational amplifier as an integrator, a feedback capacitor must be selected with a dielectric leakage current that is less than the bias current of the amplifier. Polystyrene and Teflon are usually the best choices for the ultimate in long-term integrating accuracy. If shorter integration times are required, the requirements on capacitor quality can accordingly be relaxed. Mylar capacitors may then prove satisfactory, as will silver–mica types if small values of capacitance, corresponding to high-speed integration, are to be used.

The choice of the type of amplifier is also governed by the length of computing time and the desired accuracy. Chopper-stabilized amplifiers are usually used for long-term integrators because of their superior long-term dc stability. FET amplifiers are used for medium-length integration because of their low bias current. Amplifiers with bipolar transistor input stages may be used in very-short-term integration, such as in signal generation (sweep generation, triangle waves, etc.).

If the finite gain and bandwidth are taken into account, their effects on the integrator response function may be evaluated. The open-loop frequency response of the amplifier is approximated by a single pole located at $1/\tau_o$, and a low-frequency gain of A_o.

FIGURE 13. Bode plots for amplifier and integrator. (Actual integrator response shown in color.)



The resulting integrator response function is

$$\frac{E_o}{E_1}(s) \approx \frac{-A_o}{\left(\frac{\tau_o}{A_o}s + 1\right)(A_oRCs + 1)}$$

if $A_o \gg 1$ and $A_oRC \gg \tau_o$. This function has two poles on the real axis—as opposed to the ideal integrator function, which has a single pole at the origin. In Fig. 13 the frequency response of this approximate integrator is compared with the response of an ideal integrator, along with an open-loop frequency response of the operational amplifier. Note that the response of the real integrator departs from the ideal response only at the extremes of frequency. At low frequencies, the departure is attributable to the finite gain of the operational amplifier; at high frequencies, to the finite amplifier bandwidth.

The transient response of the integrator (Fig. 14) is studied by calculating the response to a step function. The response of an ideal integrator to a step function $-E/s$ would be a linear ramp voltage increasing to infinity. The step response of the practical integrator is a close approximation of this ramp throughout most of the signal range:

$$e_o(t) = A_oE\left[1 - \frac{\epsilon^{-t/A_oRC}}{1 - (\tau_o/A_oRC)} + \frac{\epsilon^{-t/\tau_o}}{(A_oRC/\tau_o) - 1}\right]$$

In order to compare the ideal and real responses it is necessary to examine the responses for very small and for very large times. For small values of time,

$$e_o(t) \approx E\left(\frac{t}{RC} + \frac{\tau_o}{RC} + \frac{\epsilon^{-t/\tau_o}}{RC/\tau_o}\right)$$

For large values of time the response is approximately

$$e_o(t) = A_oE(1 - \epsilon^{-t/A_oRC})$$

For small values of time the principal error effect is caused by the finite bandwidth, which causes a time-lag error in the actual response. For large values of time the output signal would approach an exponential with time constant A_oRC and final value A_oE. For accurate computation, the integration should be terminated at time much less than A_oRC and output amplitude much less than A_oE.

Figure 15 illustrates the switching techniques used to initiate and terminate the period of the integration. This integrator circuit has three modes. The first of these is RESET, in which the initial conditions are established by placing an initial charge on the capacitor. This is done by closing switch S_1 to allow the output voltage to rise to the negative of V_{IC}. If switch S_1 is then opened and S_2 is closed, the circuit begins integration of the input signal e_1, beginning at the value $-V_{IC}$. This is the second or INTEGRATE mode. If both switches are held open, the output voltage will hold its latest value and will not respond to input or initial condition voltages. During this HOLD mode, the only discharge of the capacitor is that resulting from the bias current of the amplifier and dielectric leakage in the capacitor. Since electronic switch modules are commonly used for the mode-control function in place of the simple switches shown, any leakage current flowing from these switches must be added to the amplifier bias current in calculating the decay of the capacitor voltage during HOLD or during the INTEGRATE mode.

Although the analog integrator is a linear device, its maximum rate of change of output signal can lead to

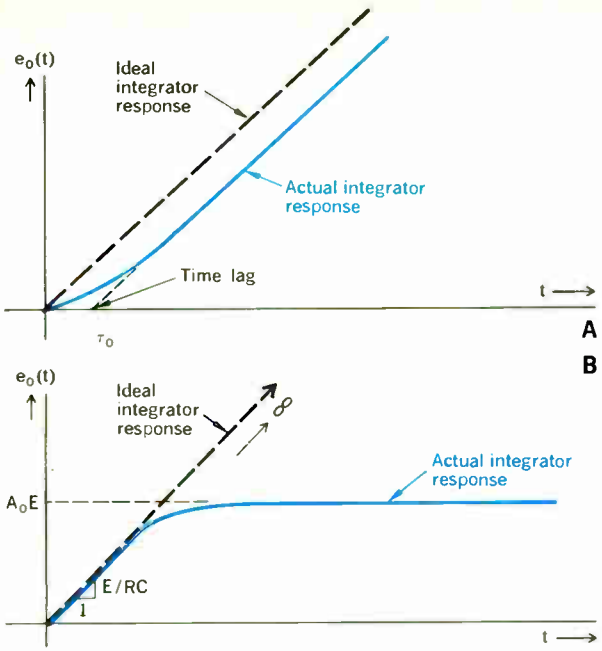


FIGURE 14. Integrator step-response curves A—For small values of time. B—For large values of time.

FIGURE 15. Three-mode integrator.

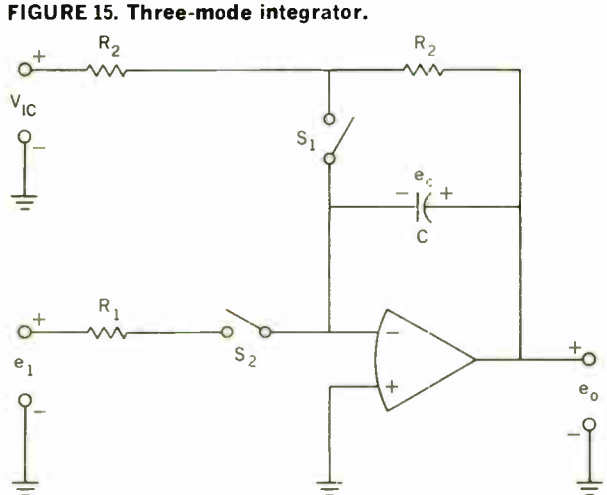
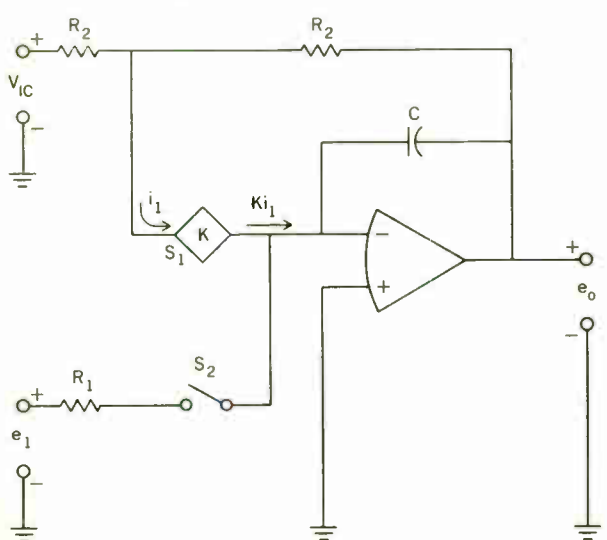


FIGURE 16. Current-amplifying switch used for reset of integrator circuit.



slew-rate distortion for signals of relatively high frequency and large amplitude. The inherent slew-rate limit of the operational amplifier places one of these limitations on the operation. However, another limitation, usually much more restrictive, is that placed on the rate of change of capacitor voltage by the output-current limits of the amplifier. The expression for this is

$$\left(\frac{de_c}{dt}\right)_{max} = \left(\frac{de_o}{dt}\right)_{max} = \frac{I_{lim}}{C}$$

where I_{lim} is the output-current limit.

The time required for the amplifier to reset to initial conditions is limited by the RC time constant of the RESET network and also by the slew rate achievable in the closed-loop circuit. If a reset switch that has a large current gain factor is used, the reset time can be considerably reduced. The use of such a switch is illustrated in Fig. 16, where the circuit is shown in the RESET state. Analysis of the circuit yields the equation for the output voltage,

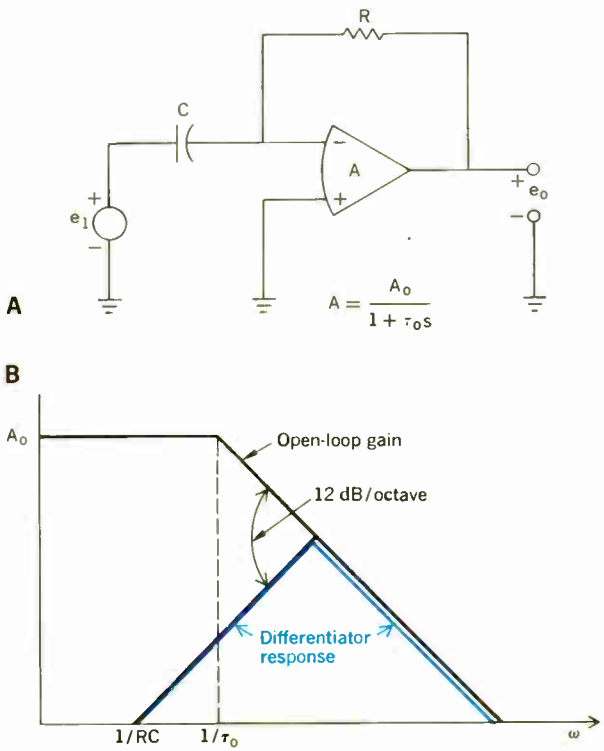
$$e_o = -e_1(1 - \epsilon^{-Kt/R_2C})$$

This is again the equation of an exponentially increasing voltage. Here, however, the time constant is R_2C/K, reduced by a factor equal to the current gain of the switch. The reset time can potentially be reduced by the factor K if the operational amplifier and switched current amplifier do not reach their current limits, thus limiting the slew rate. Maximum current is required at $t = 0$, the initiation of the RESET mode.

Differentiators[1,2,7]

By interchanging the resistor and capacitor of an integrator circuit we obtain the inverse function, differentiation. However, as will be shown, the differentiator

FIGURE 17. Differentiator using operational amplifier. A—Circuit. B—Bode plots.



circuit, shown in Fig. 17(A), has some troublesome properties. If the usual single-pole open-loop gain function is assumed for the amplifier, the transfer function of the differentiator circuit may be reduced to

$$\frac{e_o}{e_i} = \frac{-RCs}{1 + \frac{1}{A_o}(\tau_o + RC)s + \frac{RC\tau_o}{A_o}s^2}$$

This transfer function has the form

$$H(s) = \frac{-H_o s}{1 + \frac{\alpha}{\omega_n}s + \frac{s^2}{\omega_n{}^2}}$$

where $$\omega_n{}^2 = \frac{A_o}{RC\tau_o}$$

and $$\alpha \text{ (damping factor)} = \sqrt{\frac{(\tau_o + RC)^2}{RC\tau_o}}\frac{1}{A_o} \ll 1$$

Thus the damping factor α is very small, indicating a lightly damped circuit response and complex poles near the $j\omega$-axis. Such a response would also be indicated by the 12-dB/octave rate of closure of the Bode plots; see Fig. 17(B). Thus the differentiator circuit, as shown, has a tendency toward instability. If the amplifier open-loop gain has an attenuation rate of greater than 6 dB/octave over a portion of its Bode plot, the circuit may well oscillate. Another problem with this differentiator circuit is its high gain at high frequencies. This allows the high-frequency components of amplifier noise to be amplified even though the signal may not have high-frequency components. Thus the high-frequency output noise may obscure the differentiated signal.

The modified differentiator circuit of Fig. 18(A) is usually preferred as a means of eliminating the problems of the simpler circuit. Two additional real poles are introduced by use of R_1 and C_F. This creates a very stable system and reduces the high-frequency noise. The poles are placed sufficiently high in frequency to prevent significant phase-shift error in the signal-frequency range. The modified frequency response is shown in Fig. 18(B).

Line-driving amplifiers

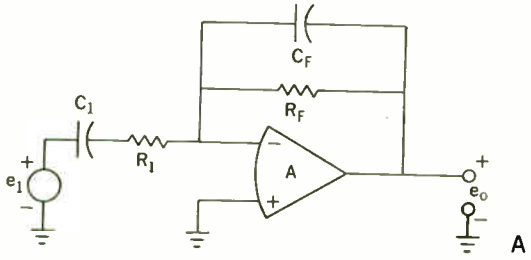
One of the primary areas of application for the operational amplifier is that of buffering between a signal source and the desired load. Usually the signal source is very limited in power, has relatively high internal impedance, and is of a low level. The load is relatively low in impedance (possibly capacitive) and requires high-level signals. Thus the amplifier must provide impedance buffering, signal scaling, and power gain. Needless to say, it must be stable under the desired conditions of loading and feedback, and must have sufficient gain and bandwidth to insure accurate response to input signals. A typical example of such an application is the line-driving amplifier.

When data signals must be transmitted over long signal lines from a remote measuring station, the line-driving amplifier is usually required. Figure 19 illustrates a simulated load of this type. The capacitance is that of a shielded cable and may be as little as a few picofarads or as much as several microfarads. If the output impedance of the amplifier is considered, the equation for effective open-loop gain $A'(s)$ becomes

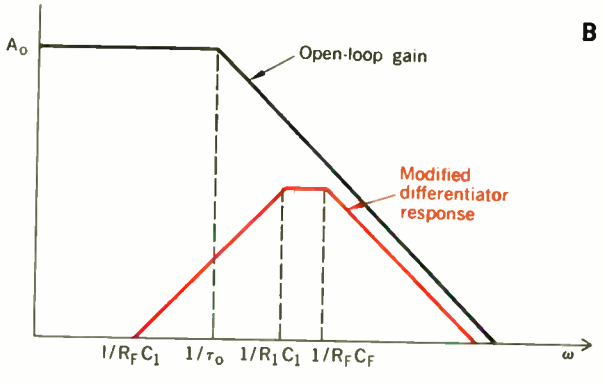
$$A'(s) = A(s)\frac{R_p}{R_p + R_o}\left(\frac{1}{1 + R_qC_Ls}\right)$$

where

FIGURE 18. A—Modified differentiator with improved noise and stability. B—Modified frequency response.



A



B

$\tau_F = R_1C_1$ $\tau_F = R_FC_F$

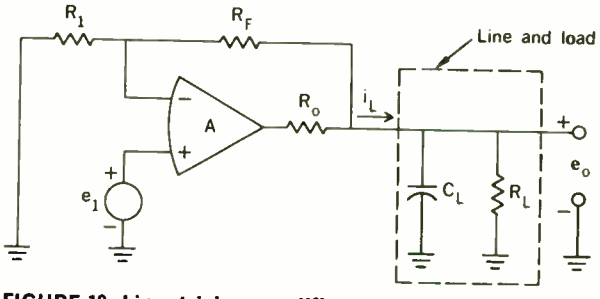
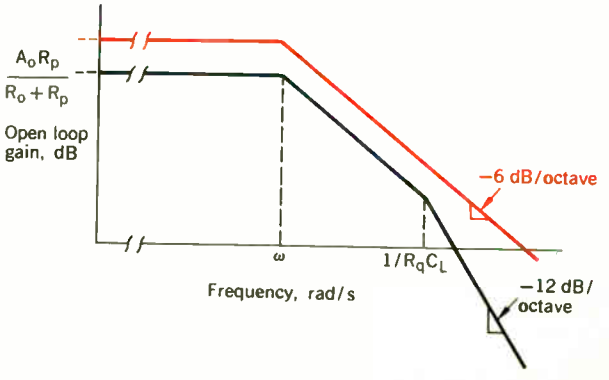


FIGURE 19. Line-driving amplifier.

FIGURE 20. Effect of loading on open-loop gain.



$$R_p = \frac{1}{\frac{1}{R_F} + \frac{1}{R_L}} \qquad R_q = \frac{1}{\frac{1}{R_F} + \frac{1}{R_L} + \frac{1}{R_o}}$$

where $A(s)$ is the unloaded open-loop gain and R_o is the dynamic output impedance of the operational amplifier. If $A(s)$ is approximated by a single-pole transfer function,

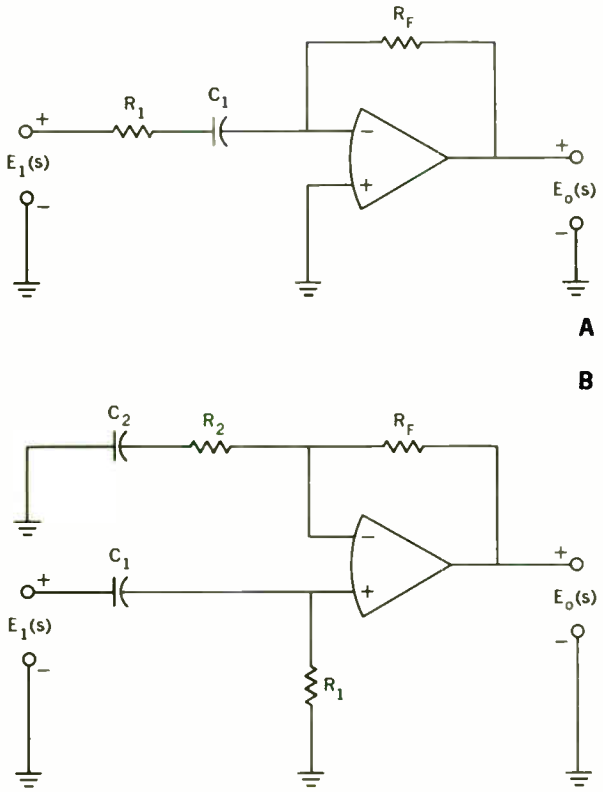
$$A(s) = \frac{A_o}{1 + (s/\omega_o)}$$

then the effective (loaded) open-loop gain becomes

$$A'(s) = \frac{R_p}{R_p + R_o}\left[\frac{A_o}{1 + (s/\omega_o)}\right]\frac{1}{1 + R_qC_Ls}$$

A Bode plot of this transfer function, for $s = j\omega$, is shown in Fig. 20, along with a plot of the unloaded open-loop gain. Note that the effect of the resistive loading is to reduce the open-loop gain, lowering the entire curve. Thus, resistive loading alone reduces the unity-gain bandwidth and will consequently reduce closed-loop bandwidth by the same factor. This bandwidth reduction factor is extremely important for fast line-drivers since the very low impedance of the line can severely degrade the bandwidth unless the operational amplifier has very low output impedance. The capacitive component of load impedance introduces another pole in the gain function at $s = -1/R_qC_L$. This causes an additional "break" in the frequency response and a rolloff of −12 dB/octave above the frequency $\omega = 1/R_qC_L$. If the closed-loop gain curve intersects this section of the effective open-loop gain curve, the amplifier will be marginally stable with unacceptable transient response.

FIGURE 21. Typical ac-coupled feedback amplifiers. A—Inverting circuit. B—Noninverting circuit.



There are a number of techniques for dealing with the problems of loading. The most satisfactory of these is to choose an amplifier with very low open-loop output impedance, or to create one by adding a power-booster stage to an available operational amplifier. This will reduce the gain and bandwidth loading factors caused by the load resistance and will increase the frequency at which the additional pole occurs. The higher in frequency this pole occurs, the more stable the closed-loop response will be. The power output stage also supplies the current necessary to meet the condition

$$(i_L)_{max} = C_L\left(\frac{de_o}{dt}\right)_{max}$$

As an example, the amplifier must be capable of supplying 63 mA to the capacitive load if C_L = 10 000 pF and the output voltage is a 10-volt sine wave at 100 kHz.

AC-coupled feedback amplifiers[1,2]

Although the operational amplifier is designed to amplify dc signals, it has a rather broad frequency response and is consequently quite useful for strictly ac signals. The feedback network can be tailored for exactly the desired passband. One of the simplest ac amplifiers is that shown in Fig. 21(A), where the closed-loop gain is given by

$$\frac{E_o}{E_i}(s) = -\left(\frac{R_F}{R_1}\right)\frac{s}{s + (1/R_1C_1)}$$

The dc gain is zero, whereas the high-frequency gain approaches $-R_F/R_1$. The lower cutoff frequency is

$$f_c = \frac{1}{2\pi R_1C_1}$$

The dc output offset voltage E_{OS} is equal to the dc input offset voltage plus the dc offset voltage generated by the input bias current flowing through R_F.

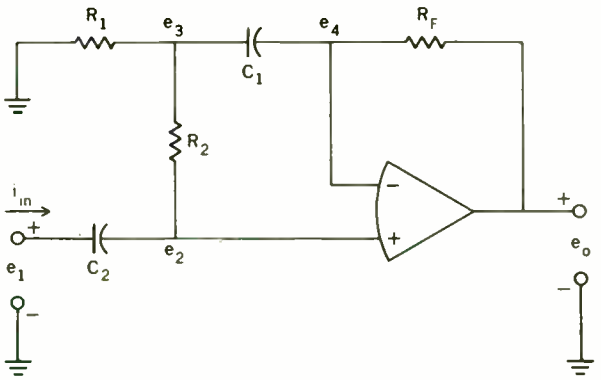
$$E_{OS} = V_{OS} \times 1.0 + I_{B1}R_F$$

A noninverting ac amplifier is shown in Fig. 21(B). The response is given by

$$\frac{E_o}{E_i}(s) = \frac{s}{s + (1/R_1C_1)}\left[\frac{(R_2 + R_F)C_2s + 1}{R_2C_2s + 1}\right]$$

Both circuits of Fig. 21 have relatively low input impedance above the cutoff frequency, determined by the

FIGURE 22. Bootstrapped ac amplifier.



resistors denoted R_1 in both cases.

The circuit of Fig. 22 is an ac amplifier whose input impedance is "bootstrapped" to a high value. Resistor R_2 provides a decoupling for dc input signals. However, for high-frequency signals the voltage across R_2 becomes very small. Consequently, very little current flows through R_2, and the effective input impedance is very high.

The analysis of the circuit is greatly simplified if it is assumed that $e_2 = e_4$ as $A \to \infty$. Then we may write the equations

$$\frac{e_1 - e_2}{X_2} = \frac{e_2 - e_3}{R_2}$$

$$\frac{e_o - e_2}{R_F} = \frac{e_2 - e_3}{X_1}$$

$$\frac{e_2 - e_3}{R_2} + \frac{e_2 - e_3}{X_1} = \frac{e_3}{R_1}$$

where $X_1 = \dfrac{1}{j\omega C_1}$ and $X_2 = \dfrac{1}{j\omega C_2}$

If these equations are solved for e_2, the input impedance may be calculated from

$$Z_{in} = \frac{e_1}{i_{in}} = \frac{e_1 X_2}{e_1 - e_2}$$

which yields

$$Z_{in} = X_2 + R_2 + R_1 + \frac{R_1 R_2}{X_1}$$

As the frequency increases, X_1 and X_2 approach zero and the input impedance becomes very large. As frequency increases still further, the open-loop gain decreases and the condition $e_2 = e_4$ is no longer enforced. The input impedance then decreases.

Differential ac amplifiers are also easily realized through the use of operational amplifiers. Two examples are shown in Fig. 23. In Fig. 23(A) a simple dc decoupling is introduced into the familiar differential dc amplifier circuit. The circuit of Fig. 23(B) provides high input impedance while decoupling dc signals in the second stage. The dc offset voltages of the first-stage amplifiers are removed by the capacitive coupling. The dc offset voltage of the second-stage amplifier is multiplied by the dc gain, 1.0.

Voltage-to-current converters[1,3]

In applications such as coil-driving and transmission of signals over long lines, it is sometimes desirable to convert a voltage to an output current. With operational amplifiers this is quite easily done. Several realizations of the voltage-to-current converter will be examined in this section.

The simplest *V*-to-*I* converters are those for floating loads. The circuits of Fig. 24 are the prime examples of this type. Figure 24(A) illustrates a simple inverting circuit. The input current is given by $i_1 = e_1/R_1$, since R_1 is terminated at the virtual ground of the summing junction. This same current flows through the feedback load impedance Z_L in the feedback loop. The current i_1 is independent of the value of Z_L. Both the signal source and the operational amplifier must be capable of supplying the desired amount of load current. The circuit of Fig. 24(B) operates in the noninverting mode and, hence, presents

FIGURE 23. Differential ac amplifiers. A—Simple one-amplifier circuit. B—High-input-impedance circuit.

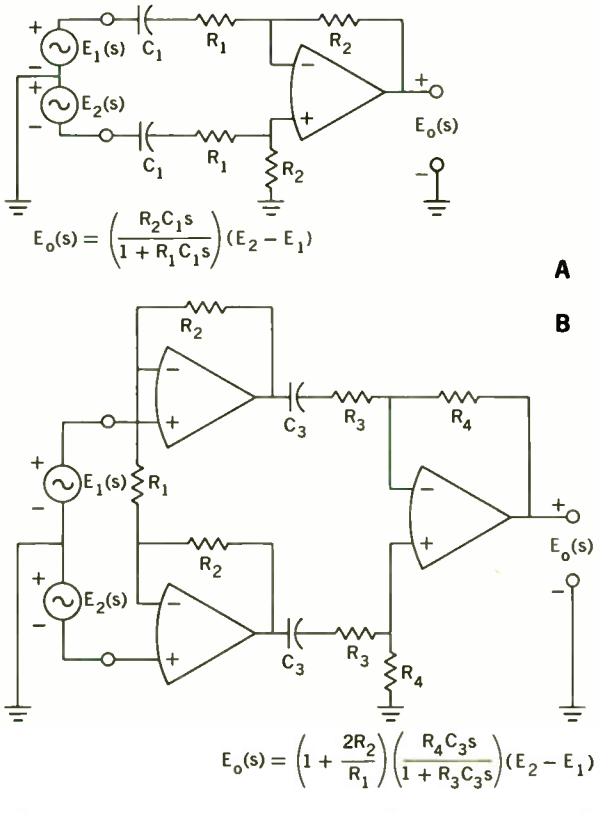
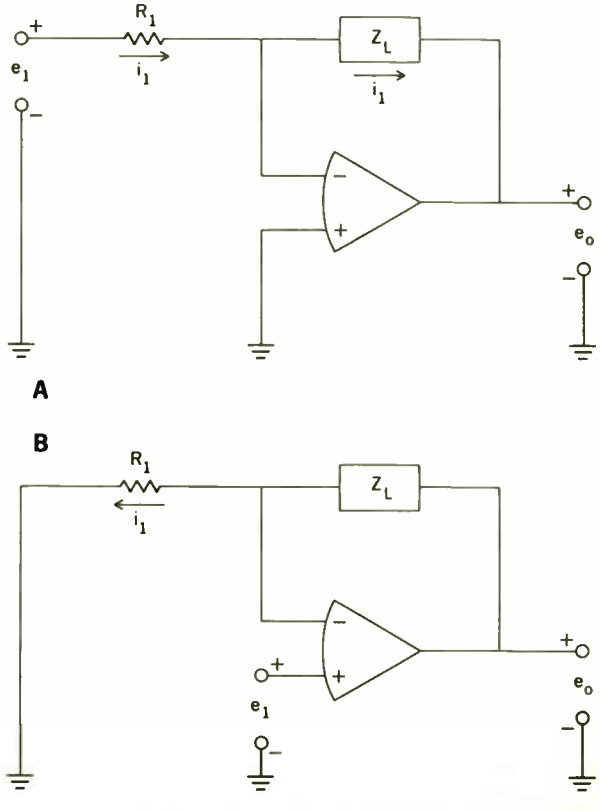


FIGURE 24. V-to-I converters, floating loads. A—Inverting amplifier type. B—Noninverting amplifier type.



a high impedance to the driving source. The current is again given by the equation $i_1 = e_1/R_1$ and, again, i_1 is the load current. Very little current, however, is required from the signal source because of the high input impedance of the noninverting amplifier.

Another V-to-I converter for a floating load is shown in Fig. 25. In this case most of the current is provided by the amplifier and only a small portion by the signal source. Analysis of the circuit yields the following equation for load current:

$$i_L = \frac{e_1}{R_1}\left(1 + \frac{R_2}{R_3}\right)$$

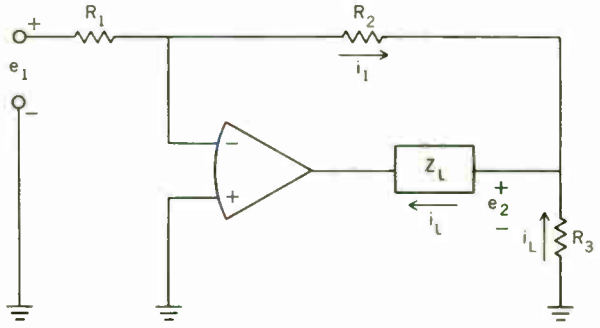


FIGURE 25. Current-amplifying circuit.

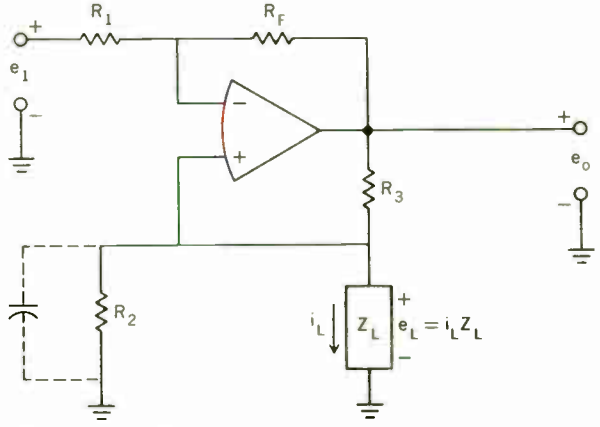
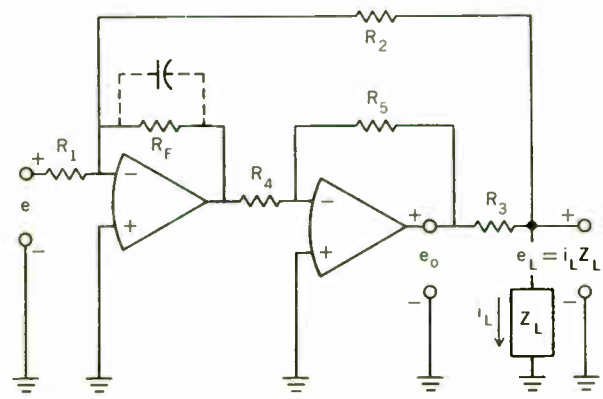


FIGURE 26. Voltage-to-current converter, grounded load.

FIGURE 27. Two-amplifier voltage-to-current converter, grounded load.



Resistor R_3 provides a convenient means for scaling the current. Resistor R_1 can be made relatively large to minimize the loading of the signal source. The amplifier must be capable of providing all of the current to the load and must also be capable of an output voltage equal to

$$(e_o)_{max} \approx (i_L)_{max}(Z_L + R_3)$$

For loads grounded on one side there are also circuits that give V-to-I conversion. The single-amplifier circuit of Fig. 26 acts as a current source controlled by e_1; i.e.,

$$i_L = -\frac{e_1}{R_2} \qquad \text{if } \frac{R_3}{R_2} = \frac{R_F}{R_1}$$

If these ratios of resistances are matched, the circuit will function as a true source of current with very high internal impedance. A mismatch of the ratios will be seen as a decreased internal impedance of the current source. Fluctuations in effective load impedance will then cause fluctuations of the output current. The operational amplifier for the circuit of Fig. 26 must have an output voltage range sufficient to provide the maximum load voltage plus the voltage drop across R_3. Normally, R_1 and R_2 will be chosen to draw small currents and R_F and R_3 will be made small to minimize voltage drops.

The circuit of Fig. 27 utilizes two inverting amplifiers to drive a current into a grounded load. This current is given by the expression

$$I_L = E_1 \frac{R_5R_F/R_4R_1}{R_3 + Z_L\left(1 + \frac{R_3}{R_2} - \frac{R_5}{R_4}\frac{R_F}{R_2}\right)}$$

If resistors are selected so that

$$1 + \frac{R_3}{R_2} = \frac{R_5R_F}{R_4R_2}$$

then

$$i_L = \frac{e_1}{R_3}\left(\frac{R_5R_F}{R_4R_1}\right)$$

In particular, if

$$R_1 = R_F = R_4 = R_5$$

then

$$i_L = \frac{e_1}{R_3} \qquad \text{and} \qquad R_2 = R_F - R_3$$

If R_1 is large, very little current is drawn from the signal source and very little flows through the feedback elements. Then the output voltage is given by

$$(e_o)_{max} \approx (I_L)_{max}(Z_L + R_3)$$

Note that, in the circuits of Figs. 26 and 27, when the load is open-circuited the positive feedback is equal to the negative feedback. This is equivalent to an open-loop condition. The stabilizing capacitors shown by dashed lines are therefore desirable to prevent excessive noise and possible oscillations. Figure 28 illustrates a modified form of the two-amplifier V-to-I converter that provides the additional feature of very high input impedance. The expression for output current as a function of input voltage is

$$i_L = \frac{e_1\frac{R_5}{R_4}\left(1 + \frac{R_F}{R_2} + \frac{R_3}{R_2}\right)}{R_3 + Z_L\left(1 + \frac{R_3}{R_2} - \frac{R_5R_F}{R_2R_4}\right)}$$

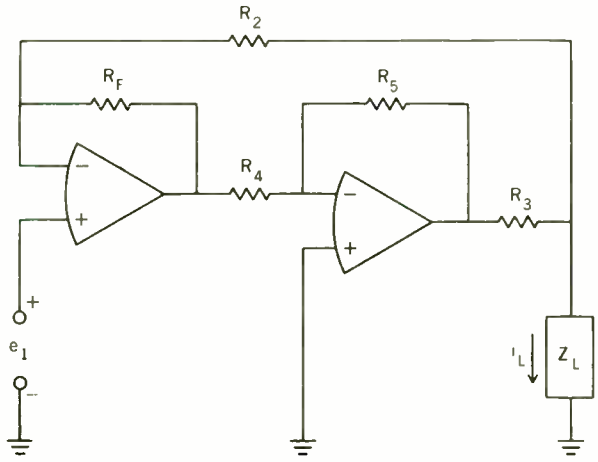
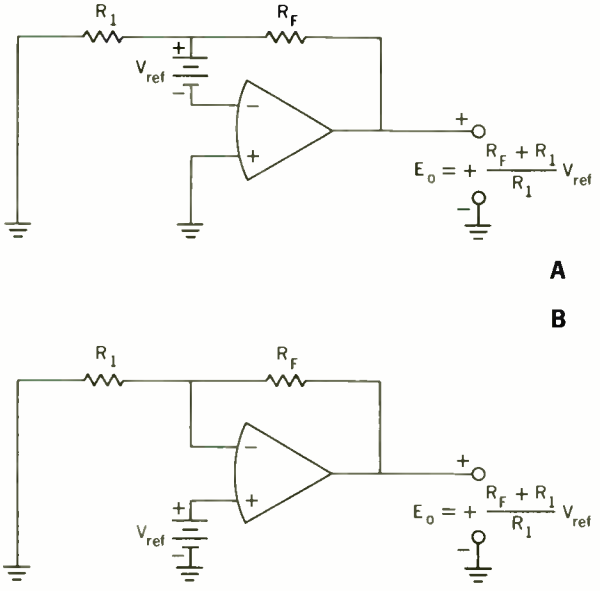


FIGURE 28. Buffered V-to-I converter, grounded load.

FIGURE 29. Reference-voltage sources. A—Single-ended circuit. B—Noninverting circuit.



If we again select resistors such that

$$1 + \frac{R_3}{R_2} = \frac{R_5 R_F}{R_2 R_4} \quad \text{and} \quad R_F = R_4 = R_5$$

then $$i_L = \frac{2e_1 R_F}{R_2 R_3} \quad \text{and} \quad R_2 = R_F - R_3$$

Reference-voltage sources and regulators[1-3]

Because of its high input impedance and easily adjustable gain, the operational amplifier may be used as a reference-voltage source with very low output impedance and substantial output-current capability. Two circuits for use with standard cells are shown in Fig. 29. In both instances the output voltage is given by

$$E_o = V_{ref}\left(1 + \frac{R_F}{R_1}\right)$$

The circuit of Fig. 29(A) can be used with single-ended amplifiers (such as chopper-stabilized types), as well as with those having a differential input. The circuit of

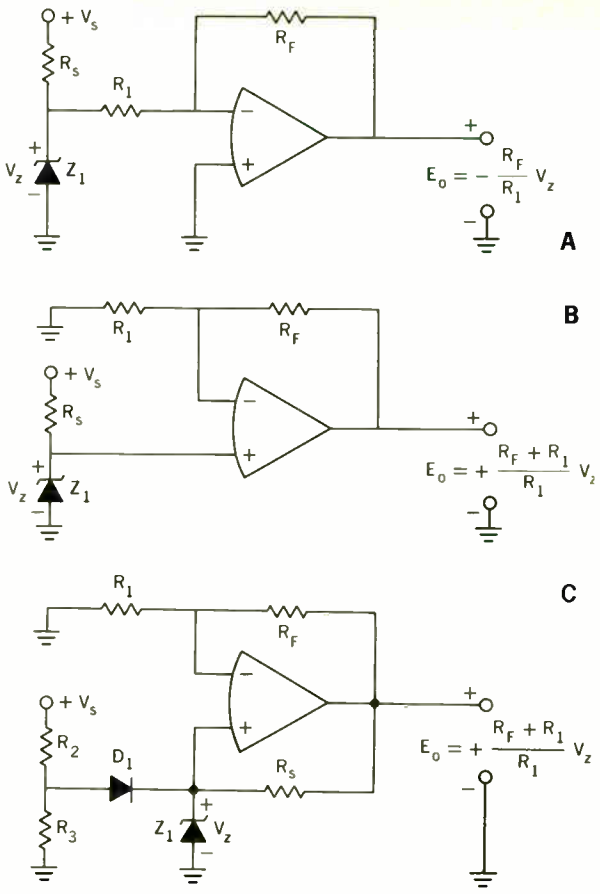


FIGURE 30. Zener reference sources. A—Inverting. B—Noninverting. C—Noninverting with regulated Zener drive.

Fig. 29(B) is used if the reference source or cell must be grounded on one side. The only current drawn from the cell is the input bias current of the amplifier plus a term given by

$$I_{in} = \frac{E_o}{AR_{in}} = \frac{V_{ref}\left(1 + \frac{R_F}{R_1}\right)}{AR_{in}}$$

where R_{in} is the differential input impedance of the operational amplifier. This component of current is negligible in comparison to bias current for most amplifiers. The reference voltage cell is, for all practical purposes, isolated from any load being driven. The effective output impedance, R_{out}, is given by

$$R_{out} = \frac{R_o}{A\beta}$$

where $$\beta = \frac{R_1}{R_1 + R_F}$$

and R_o is the open-loop output impedance.

The load regulation is therefore given by

$$\text{Regulation (percent)} = \frac{R_o}{A\beta R_L} \times 100$$

where R_L is the minimum load impedance.

Similar circuits for use with Zener diodes are shown in Fig. 30(A) and (B). The loading conditions on the Zener diodes are constant and the load regulation is the

same as derived for the circuits of Fig. 29. Regulation with respect to the input voltage V_s depends upon the dynamic resistance of the reference Zener diode Z_1. The circuit of Fig. 30(C) further reduces this regulation due to input voltage by providing the output reference voltage as the source for the Zener diode current. The dc voltage V_s now functions only as a "start-up" voltage through the network of R_2, R_3, and D_1.

Voltage regulators

Any one of the voltage references described in the preceding section may be considered a voltage regulator, with extremely tight regulation characteristics. Where higher ouput currents are required, a power booster can be added, inside the feedback loop. However, in speaking of voltage regulators, it is more usual to consider operation from a single source of unregulated dc voltage, rather than the dual supplies tacitly assumed in the reference-voltage circuits. Figure 31 shows such a regulator. The amplifier, which normally operates on dual power supplies of opposite polarity, is biased for operation on a single unregulated power supply. The negative supply terminal is grounded and the noninverting input is biased at the Zener voltage. The Zener diode Z_1 operates at constant load current, since the output current is provided by the transistor Q_1. If the amplifier has a minimum (balanced) supply rating of $\pm V_m$, then V_s must be larger than $2V_m$. Similarly, if $\pm V_M$ is the maximum (balanced) supply rating, V_s must not exceed $2V_M$. The amplifier will saturate as the output voltage approaches either supply voltage. This determines the limit on output, whereas the common-mode voltage range sets the lower limit on Zener voltage.

Although the amplifier may have an internal current limit, the resistor R_p is required to protect against short circuit in this type of regulator. This is because a short circuit to ground is equivalent to a short circuit to the negative supply. This causes a power dissipation equal to twice that of a short circuit to ground when operating on balanced dual supplies. Thus the internal protection may not be sufficient. The value of R_p should be chosen to limit the amplifier short-circuit current to approximately half the internal current-limit value when the output is at positive saturation voltage. The resistor R_s provides current-limiting to protect Q_1.

The load regulation of this type of regulator can exceed 0.01 percent, since the effective output impedance is very low. The line regulation is increased beyond that of the Zener by using the output voltage as excitation for the Zener.

Current amplifiers

Current amplifiers, or current-to-voltage converters, are realized very simply by the use of operational amplifiers. An ideal current source has infinite output impedance and an output current that is independent of load. Photocells and photomultiplier tubes are basically current sources with an output impedance that is finite but very large. For small load impedances, the output impedance may be considered infinite.

The current-to-voltage converter of Fig. 32 presents almost zero load impedance to ground because the inverting input appears as a virtual ground. The input current, however, flows through the feedback resistor, generating an output voltage

$$e_o = -i_s R_F$$

The actual input impedance Z_{in} of the current-to-voltage converter, taking into account the finite gain A and differential (open-loop) input impedance Z_{id}, is

$$Z_{in} = \frac{Z_{id}}{1 + \frac{Z_{id}}{R_F}(1 + A)} \approx \frac{R_F}{1 + A}$$

The lower limit on measurement of current input is determined by the bias current of the inverting input. For greatest resolution, FET or varactor bridge amplifiers are usually employed.

The gain of the amplifier for dc offset voltage and noise voltage is given by

$$\frac{R_F + R_S}{R_S} \approx 1.0 \qquad \text{since } R_s \gg R_F$$

Although errors resulting from these parameters are very small, current noise can be a factor because of the very large impedances. Since most such measuring circuits are used for very-low-frequency signals, it is usual to parallel R_F with a capacitor C_F to reduce the high-frequency current noise. Output impedance of the current-to-voltage converter is very low because of the nearly 100 percent feedback.

Charge amplifiers

Some transducers, such as capacitance microphones and some types of accelerometers, operate on the principle of conversion of the measurement variable into an equivalent charge. The equivalent circuit of such a trans-

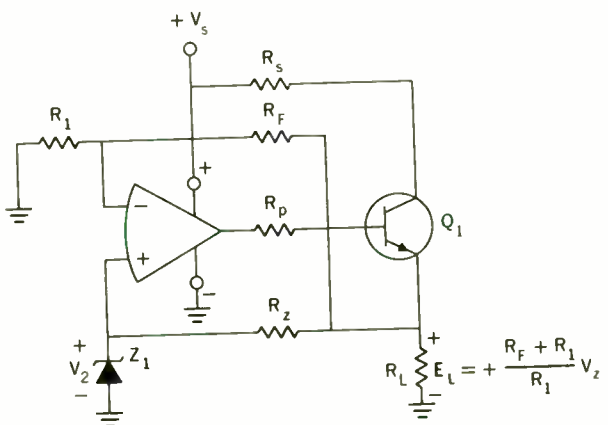
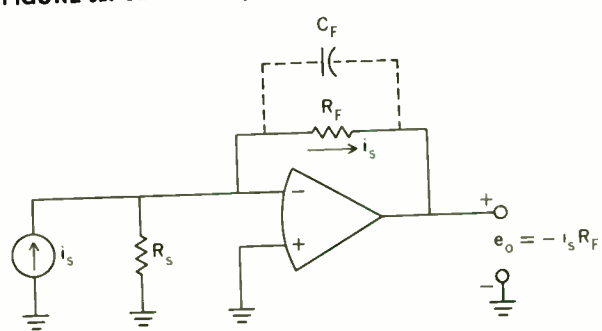


FIGURE 31. Voltage regulator.

FIGURE 32. Current amplifier (I-to-V converter).



ducer may be represented by a battery and capacitor in series, as shown in Fig. 33(A). As the capacitance varies, the charge also changes, according to the equation

$$\Delta q = \Delta C_1 E$$

When the transducer is connected to the inverting input of an operational amplifier, as in Fig. 33(A), this charge flows into the feedback capacitor C_F. The resultant change in charge on C_F generates an output voltage

$$e_o = -\Delta C_1 \frac{E}{C_F}$$

Since the operational amplifier requires a dc path from each input to common (for bias-current flow), it is necessary to insert the resistor R_F. In the absence of this resistor, the capacitors will build up a dc charge until the output voltage reaches saturation. This resistor limits the lower cutoff frequency of the charge amplifier. For stabilization purposes, and sometimes for protection of the amplifier input stage, it is also desirable to insert the series resistor R_1. This resistor limits the upper response frequency as shown in Fig. 33(B).

The gain, or sensitivity, of the charge amplifier in its passband is given by

$$\frac{e_o}{\Delta C_1} = -\frac{E}{C_F}$$

and can be varied only by changes in C_F. It is usually desirable to use a small value of C_F consistent with the desired frequency response and a reasonable value of R_F. Because of their high input impedance, low bias current, and wide bandwidth, FET amplifiers are usually the first choice for charge amplification.

Another common form of charge amplifier is shown in Fig. 34. Here the amplifier operates as a noninverting buffer with gain. Charge flows into, and out of, capacitor C_2 as the capacitance of the transducer varies. Once again these capacitance variations are converted into voltage variations at the amplifier output. An amplifier with an FET input stage is usually required in this circuit; it also minimizes the bias and noise currents. Resistor R_B provides the dc path for this bias current and limits the low-frequency response of the circuit.

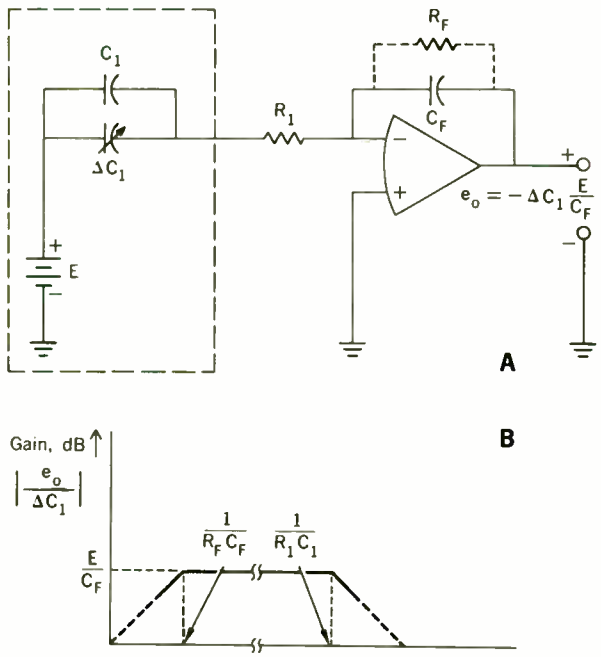
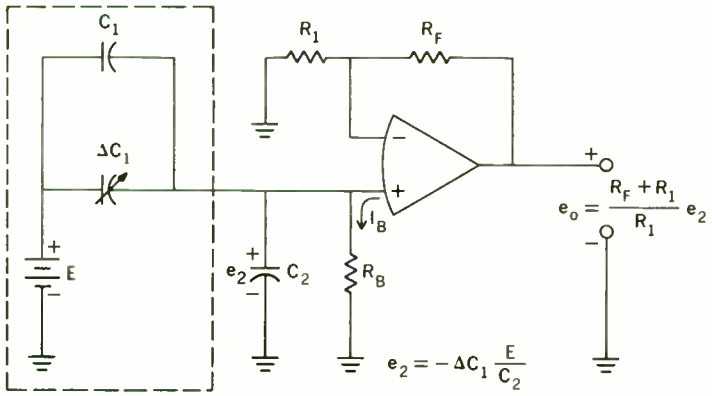


FIGURE 33. Charge-amplifier operation. A—Typical circuit. B—Frequency response.

FIGURE 34. Charge-amplifier circuit in which amplifier operates as a noninverting buffer with gain.



REFERENCES

1. *Applications Manual for Operational Amplifiers*, Philbrick/Nexus Research, 1965.
2. *Handbook of Operational Amplifier Applications*, Burr-Brown Research Corp., Tucson, Ariz., 1961 (out of print).
3. *Handbook and Catalog of Operational Amplifiers*, Burr-Brown Research Corp., Tucson, Ariz., 1969 (out of print).
4. Korn, G. A., and Korn, T. M., *Electronic Analog and Hybrid Computers*. New York: McGraw-Hill, 1964.
5. Miura, T., *et al.*, "On computing errors of an integrator," *Proc. 2nd AICA Conf.*, Presses Académiques Européenes, Brussels, 1958.
6. Tobey, G., "Analog instrumentation," *Instr. Control Syst.*, Jan. 1969.
7. Diamantides, N. D., "Improved electronic differentiator," *Electronics*, July 27, 1962.
8. Korn, G. A., "Exact design equations for operational amplifiers with four-terminal computing networks," *IRE Trans. Electronic Computers*, vol. EC-11, pp. 82–83, Feb. 1962.

Larry L. Schick (M) received the B.S.E.E. degree with honors from Kansas State University in 1953 and the M.S.E.E. degree from the University of Arizona in 1968. From 1963 to 1965 he served as project officer on engineering tests for various pieces of communication equipment at the Army Electronic Proving Ground, Fort Huachuca, Ariz. In 1966 he joined Burr-Brown Research Corp., Tucson, Ariz., as a regional applications engineer. His duties included presenting seminars, writing articles, and assisting Burr-Brown's customers in the most efficient use of operational amplifiers and other related modules, such as multipliers and A/D and D/A converters. In his present position of Western Area sales manager he has the responsibility for directing the efforts of Burr-Brown's regional applications engineers and engineering representatives in the western half of the United States. Mr. Schick is a member of Eta Kappa Nu and is a registered professional engineer in the state of Arizona.

Introduction to radio and radio antennas

The concepts of wave transmission can be more readily grasped if the approach is based upon an understanding of the fundamental circuit elements: inductors and capacitors

Harald T. Friis *Rumson, N.J.*

Electromagnetic wave propagation is a simple and beautiful phenomenon, but its rigorous mathematical derivation from Maxwell's equations is formidable. This article makes many of the concepts visible and plausible using only mathematics available to college freshmen and taught today in most high schools. Those who are familiar with the subject will appreciate the simple and direct way in which Dr. Friis' transmission formula is derived. This derivation is new and is published here for the first time.

—David DeWitt, Editor

Transmission lines guide telephone signals, and also power, from one place to another. Although such guidance does not exist for radio signals, it is illustrative to use the transmission of waves between two parallel conductors as an introduction to radio.

A transmission line has a capacitance C and an inductance L, per unit length, and in order to derive the transmission-line equations we must know the fundamental properties of the important circuit elements: inductors and capacitors.

Inductors

Figure 1 shows a long coil of wire with direct current i. The resistance of the wire is assumed to be negligible. By means of a magnetic needle indicator it has been found that the magnetic field H inside the coil is constant. Physicists also found for a closed path around currents that the sum of the magnetic fields along the path multiplied by the path lengths is equal to the total current through the area enclosed by the path. Path $ABCDA$ in Fig. 1 shows such a path. The magnetic field is zero along this path except along AD, and the total current through the area of the path is $i \times n$. Thus,

$$Hl = in$$

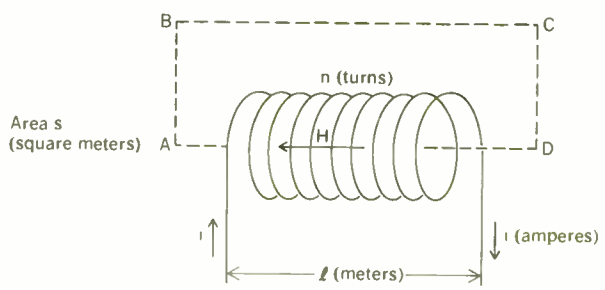
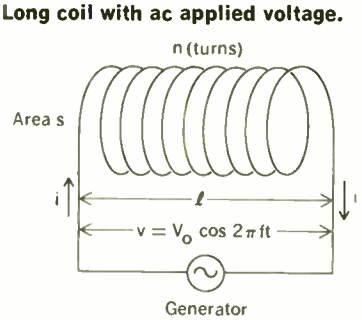


FIGURE 1. Long coil of wire with direct-current flow.

FIGURE 2. Long coil with ac applied voltage.



or, by definition,

$$H = \frac{in}{l} \text{ amperes per meter}$$

In Fig. 2 an ac voltage $v = V_0 \cos 2\pi ft$ is applied to the coil; f is the frequency in hertz and t is the time in seconds. (See Fig. 3 for variation of v vs. $2\pi ft$.) Faraday found that the voltage induced in one turn by i is proportional to the area s of the turn and the rate of change of i or H. Hence,

$$v \text{ (one turn)} = \mu s \frac{dH}{dt} = \mu s \frac{n}{l} \frac{di}{dt}$$

where the proportionality factor μ is the permeability of the medium inside the coil (for air, $\mu = 4\pi \times 10^{-7}$ H/m). The voltages in the separate turns are aiding. Therefore, the voltage induced across the coil is

Editor's note: Last month an autobiographical memoir by Dr. Friis, *Seventy-Five Years in an Exciting World*, was published by San Francisco Press, 555 Howard St., San Francisco, Calif.; it also contains "The Wisdom of Harald T. Friis," based on a famous speech (annotated by John R. Pierce) in which the author distills a lifetime's hard-won research experience.

$$v = \mu s \frac{n^2}{l}\frac{di}{dt} = L\frac{di}{dt} \text{ volts} \qquad (1)$$

$L = \mu s(n^2/l)$ is called the inductance of the coil and is measured in henrys.

For $v = V_0 \cos 2\pi ft$, Eq. (1) may be solved by integration to obtain

$$i = \frac{V_0}{2\pi fL} \sin 2\pi ft \text{ amperes} \qquad (2)$$

The values of i and v are plotted versus $2\pi ft$ in Fig. 3.

Capacitors

Figure 4 shows the plates of a capacitor. Electrostatic experiments showed that the ratio of total charge q on each plate to the voltage v between the plates is constant:

$$\frac{q}{v} = \text{constant} = C \qquad (3)$$

C is proportional to the ratio of the area s (square meters) and spacing d (meters) of the plates:

$$C = \epsilon \frac{s}{d} \text{ farads} \qquad (3a)$$

where ϵ, the proportionality factor, is the dielectric constant of the medium between the plates. For free space $\epsilon = (1/36\pi)10^{-9}$ F/m.

In Fig. 5 a voltage $v = V_0 \cos 2\pi ft$ is applied to the plates. The current i, called the displacement current, is equal to the rate of change of q. Equation (3) gives

$$i = \frac{dq}{dt} = C\frac{dv}{dt} = -2\pi fCV_0 \sin 2\pi ft \qquad (4)$$

The values of i and v are plotted versus $2\pi ft$ in Fig. 6.

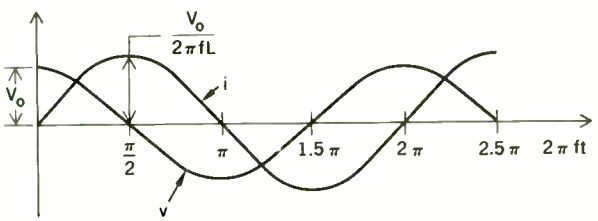


FIGURE 3. Voltage (v) and current (i) vs. $2\pi ft$ for inductive circuit of Fig. 2.

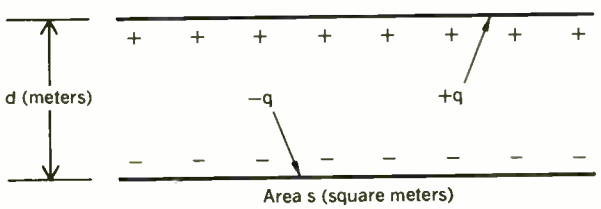


FIGURE 4. Basic capacitor configuration.

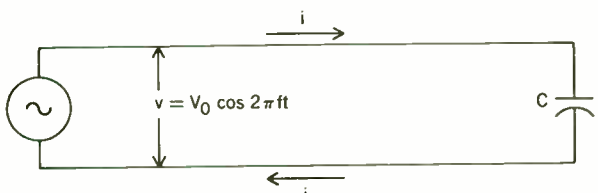
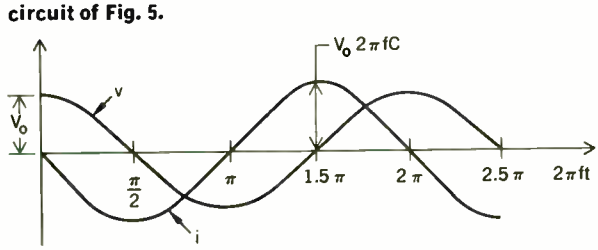


FIGURE 5. Capacitor with ac applied voltage.

FIGURE 6. Voltage (v) and current (i) vs. $2\pi ft$ for capacitive circuit of Fig. 5.



Transmission lines

Figure 7(A) shows a generator that feeds a long, loss-free uniform line with an inductance of L henries and capacitance of C farads per meter and Fig. 7(B) shows the equivalent lumped-circuit line.

The current i_x through inductor $L\,dx$ decreases the voltage from AB to CD. Equation (1) gives

$$dv_x = -L\,dx\frac{di_x}{dt}$$

or

$$\frac{dv_x}{dx} = -L\frac{di_x}{dt} \qquad (5)$$

Differentiating with respect to t gives

$$\frac{d^2v_x}{dx\,dt} = -L\frac{d^2i_x}{dt^2} \qquad (5a)$$

The displacement current through capacitor $C\,dx$ decreases the current from A to C. Equation (4) gives

$$di_x = -C\,dx\frac{dv_x}{dt}$$

or

$$\frac{di_x}{dx} = -C\frac{dv_x}{dt} \qquad (6)$$

Differentiating with respect to x gives

$$\frac{d^2i_x}{dx^2} = -C\frac{d^2v_x}{dx\,dt} \qquad (6a)$$

Substituting the value for $d^2v_x/dx\,dt$ given by Eq. (5a) gives

$$\frac{d^2i_x}{dx^2} = LC\frac{d^2i_x}{dt^2} \qquad (7)$$

Similarly, we obtain

$$\frac{d^2v_x}{dx^2} = LC\frac{d^2v_x}{dt^2} \qquad (8)$$

Mathematicians can solve for i_x and v_x in Eqs. (7) and (8) and get, as one correct solution,

$$i_x = I_0 \sin(2\pi ft - 2\pi fx\sqrt{LC}) \qquad (9)$$

$$v_x = V_0 \sin(2\pi ft - 2\pi fx\sqrt{LC}) \qquad (10)$$

We can verify that these solutions are correct by differentiating i_x in Eq. (9) twice with respect to x and twice with respect to t:

$$\frac{di_x}{dx} = -I_0 \cos(2\pi ft - 2\pi fx\sqrt{LC}) \times 2\pi f\sqrt{LC}$$

and

$$\frac{d^2i_x}{dx^2} = -I_0 \sin(2\pi ft - 2\pi fx\sqrt{LC}) \times (2\pi f\sqrt{LC})^2 \quad (11)$$

Also,

$$\frac{di_x}{dt} = I_0 \cos(2\pi ft - 2\pi fx\sqrt{LC}) \times 2\pi ft$$

and

$$\frac{d^2i_x}{dt^2} = -I_0 \sin(2\pi ft - 2\pi fx\sqrt{LC}) \times (2\pi ft)^2 \quad (12)$$

Equations (11) and (12) give

$$\frac{d^2i_x}{dx^2} = LC\frac{d^2i_x}{dt^2} \quad (13)$$

Equations (7) and (13) are identical. Therefore, Eq. (9) gives a correct solution of (7) and, similarly, (10) gives a correct solution of (8).

We substitute from Eqs. (9) and (10) in (5):

$$-2\pi f\sqrt{LC}V_0 \cos(2\pi ft - 2\pi fx\sqrt{LC}) = -LI_0 2\pi f \cos(2\pi ft - 2\pi fx\sqrt{LC})$$

That is,

$$\sqrt{LC}V_0 = LI_0$$

or

$$V_0 = I_0\sqrt{\frac{L}{C}}$$

From Eqs. (9) and (10),

$$v_x = i\sqrt{\frac{L}{C}}$$

or the impedance of the line is

$$Z = \frac{v_x}{i_x} = \frac{V_0}{I_0} = \sqrt{\frac{L}{C}} \text{ ohms} \quad (13a)$$

Values of v_x, as given by Eq. (10), versus x are plotted in Fig. 8 for $t = 0, t = 1/4f, t = 1/2f, \ldots$.

Note that the first voltage maximum moves with time from A to B to C to D to E to F with a speed

$$c = \frac{1}{2f\sqrt{LC}}\Big/\frac{1}{2f} = \frac{1}{\sqrt{LC}} \text{ m/s}$$

or the input generator causes a traveling wave on the line.

The separation between voltage maximums is called the wavelength λ:

$$\lambda = \frac{1}{f\sqrt{LC}} = \frac{c}{f} \text{ meters}$$

Equation (10) shows that the difference in phase ϕ of the voltage at distance x and distance $x + d$ is

$$\phi = 2\pi f(x + d)\sqrt{LC} - 2\pi fx\sqrt{LC} = 2\pi fd\sqrt{LC} = 2\pi\frac{d}{\lambda} \quad (14)$$

Figure 9 shows a uniform strip transmission line in which the edge effect is neglected. The transmission-line theory gave line impedance $z = \sqrt{L/C}$ ohms.

In this line the inductance L per meter is the inductance of a single-turn coil with area $s = 1h$ and length $l = a$. Equation (1) gives

$$L = \mu h\frac{1}{a}$$

Similarly, Eq. (3a) gives

$$C = \epsilon\frac{a}{h}$$

The speed of propagation c is given by

$$c = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\mu\epsilon}}$$

In free space,

$$\mu = 4\pi \times 10^{-7} \quad \text{and} \quad \epsilon = \frac{1}{36\pi} \times 10^{-9}$$

so $c = 3 \times 10^8$ m/s, which is the speed of light and also the propagation speed for transmission lines constructed

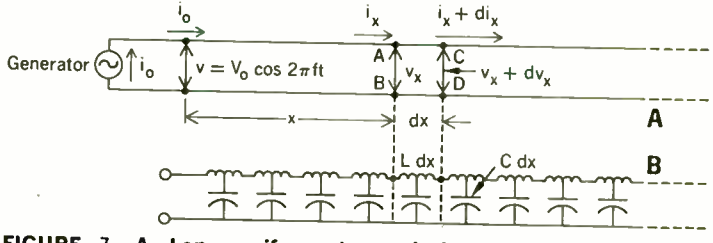


FIGURE 7. A—Long uniform transmission line. B—Equivalent lumped-circuit line.

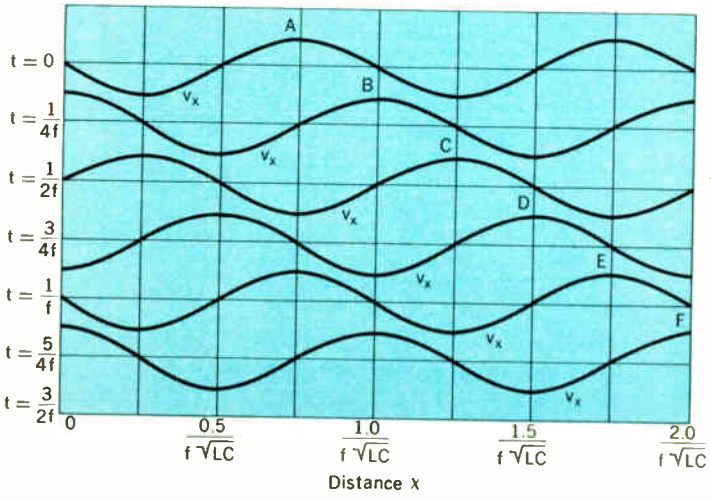
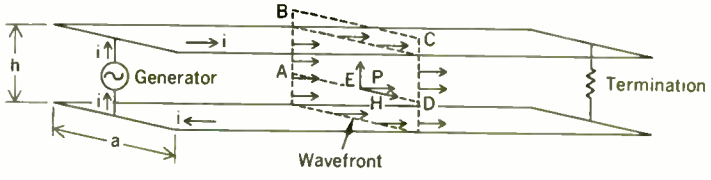


FIGURE 8. Line voltage v_x vs. distance x along the line.

FIGURE 9. Uniform strip transmission line with traveling wave.



with negligible amounts of insulation and magnetic material.

The line impedance z is given by

$$z = \frac{h}{a}\sqrt{\frac{\mu}{\epsilon}} = 120\pi\,\frac{h}{a}$$

For $h = a = 1$ and neglecting edge effects, we get Schelkunoff's formula for the impedance of free space:

$$z_{\text{free space}} = 120\pi \text{ ohms}$$

A sheet with 120π ohms between two opposite edges of a square of the sheet is called a resistance sheet with 120π ohms per square. It matches waves in free space and, if inserted between the ends of the line, makes a good termination; that is, it causes no reflected waves for the traveling wave on the line when $h < \lambda$. (A conducting sheet $\lambda/4$ behind the resistance sheet is required when h is larger.)

The voltage across the line is hE volts. As shown in Fig. 9, E is the maximum value of the electric field in the space between the strips. It is measured in volts per meter. The power absorbed by a circuit of impedance Z with an input voltage E is $\frac{1}{2}E^2/Z$. The impedance of this transmission line is $Z = 120\pi h/a$. The power flow in the line is, therefore,

$$\frac{\frac{1}{2}(hE)^2}{z} = \frac{\frac{1}{2}E^2}{120\pi}\,ha \text{ watts}$$

The power flow per unit area or the power intensity in the wave between the strip is then

$$\frac{\frac{1}{2}E^2}{120\pi} \text{ watts (same for radio waves)} \qquad (14a)$$

Note that all the power flows in the air between the strip conductors. This is true for all transmission lines. The conductors act only as guides for the waves.

Figure 9 also shows the magnetic field vector H. Its magnitude is found by choosing a path $ABCDA$ and adding the field times path length as explained in the first section on inductors. The current through the areas of this path is i and the field is zero except along path AD. Hence,

$$Ha = i$$

Equation (13a) gives

$$i = \frac{v}{\sqrt{L/C}} = \frac{Eh}{120\pi h/a} = \frac{Ea}{120\pi}$$

Therefore,

$$H = \frac{i}{a} = \frac{E}{120\pi}$$

Figure 9 shows the power intensity vector P, also called the Poynting vector, from (14a),

$$P = \frac{\frac{1}{2}E^2}{120\pi} = \frac{1}{2}EH \text{ watts per square meter}$$

Recapitulating what happens in the idealized line illustrated in Fig. 9, it has been shown (1) that the available power from the generator has been changed to a plane electromagnetic wave that moves with the speed of light toward the termination of the line, (2) that the power flow intensity is constant over the cross section of the wave and equal to $\frac{1}{2}EH$ watt where E is the electric field and H the magnetic field of the wave, and (3) that the impedance of the line is $120\pi h/a$ ohms and the free-space medium between the strips is 120π ohms.

The strip-line dimensions of width a and separation h may be increased gradually as shown in Fig. 10. The traveling wave will continue through this expanded section and produce a large radiating wavefront at the opening $ABCD$. We have now a radio transmitting antenna (Schelkunoff's paddle antenna). Alternatively, the strip line may feed a paraboloid, as shown in Fig. 11. We have, in other words, illustrated the passing of guided waves to radio waves; we shall next discuss some important properties of radio antennas and radio transmission.

Effective area of an antenna. The effective area A_{eff} of an antenna is defined as shown in Fig. 12, for the case of receiving, as the available power P_R from the antenna

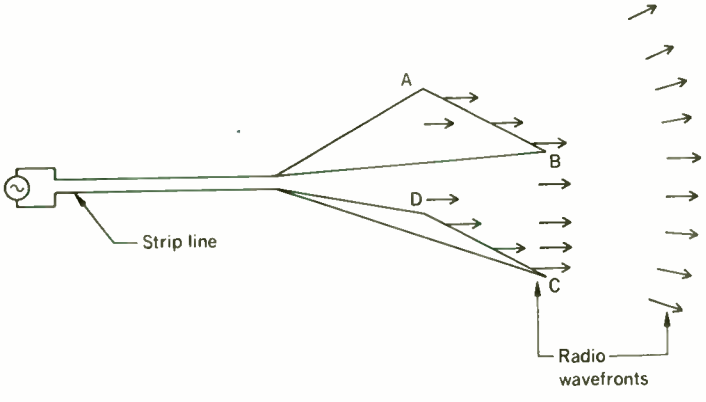


FIGURE 10. Strip line expanded.

FIGURE 11. Strip line feeding a paraboloid.

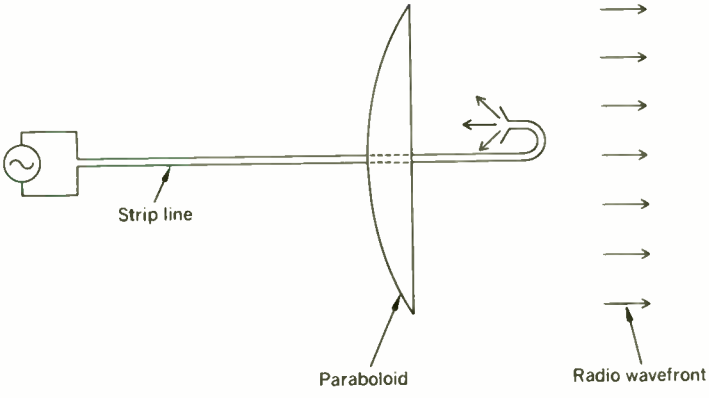
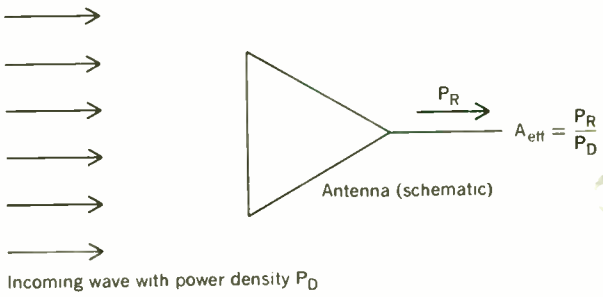


FIGURE 12. The effective area of an antenna.



divided by the power density P_D of the incoming wave:

$$A_{\text{eff}} = \frac{P_R}{P_D}$$

If the illumination of the antenna when used for transmitting is uniform (that is, if the electromagnetic field is constant and in phase over the plane of the aperture and zero outside), the effective area is by reciprocity equal to the geometrical area of the aperture.

Examples: The effective area of the paddle antenna shown in Fig. 10 is approximately equal to the opening area $ABCD$ and the effective area of the paraboloid antenna shown in Fig. 11 is about two thirds of the opening area of the paraboloid. The effective area of a small loss-free dipole, which will be calculated later, is $(\lambda/2) \times (\lambda/4)$. As usual, all linear dimensions are in meters.

Propagation loss between two antennas in free space. That light propagates as a wave was suggested some 300 years ago by Huygens, and resulted in Huygens' principle: "Every part of a wavefront can be regarded as a source of disturbance that emits a spherical wavelet." Fresnel made his great contribution some 150 years later by realizing that the relative phase ϕ between two wavelets can be calculated from their path difference d and wavelength λ [compare with Eq. (14)]:

$$\phi = 2\pi \frac{d}{\lambda}$$

Figure 13 shows how Fresnel graphically found the field due to the wavelets in a plane wave at a distant point. Figure 13(B), a side view of the wave, indicates the wavelets on the wavefront that radiate spherically toward the distant point P. Figure 13(A) shows how the wavefront is divided into equal-area rings 1, 2, 3, $\cdots$. The Huygens sources in each ring have the same distance to point P and are therefore in phase at P. Each of the rings has the same area or the same number of wavelets and therefore produces the same fields E_1, E_2, E_3, $\cdots$ at P when the angle ϕ is small.

Figure 13(C) shows how two parallel fields

$$E_1 = E \cos (2\pi ft - \phi_1)$$

and

$$E_2 = E \cos (2\pi ft - \phi_2)$$

can be added graphically by plotting $OA = E$ first and then $AB = E$ so that the angle between AB and the direction of OA is $\phi_2 - \phi_1$. The sum of E_1 and E_2 is then OB. That this is correct can be shown as follows:

$$E_1 + E_2 = E \cos (2\pi ft - \phi_1) + E \cos (2\pi ft - \phi_2)$$
$$= E[\cos (2\pi ft - \phi_1) + \cos (2\pi ft - \phi_2)]$$

The formula for the addition of two cosine functions gives

$$E_1 + E_2 = 2E \cos \left(2\pi ft - \frac{\phi_1 + \phi_2}{2}\right) \cos \frac{\phi_2 - \phi_1}{2}$$

or the amplitude of $E_1 + E_2$ is

$$2E \cos \frac{\phi_2 - \phi}{2}$$

and this is the length of OB in Fig. 13(C).

The fields at point P from the wavelets can now be added graphically, as shown in Fig. 13(D). Starting at point O, the field E_1 is plotted first with arbitrary length and direction and it is followed by E_2 so that E_1 and E_2 form an angle

$$\Delta\phi = 2\pi \frac{d_2 - d}{\lambda}$$

where d_2 is the distance from the sources in ring 2 and d the distance from the sources in the center element 1. The fields E_3, E_4, $\cdots$ are plotted by continuing in the same manner. The small phase angles $\Delta\phi$ are alike when angle ϕ is small and the ends of the fields therefore follow a circle. This circle turns into a spiral when angle ϕ increases and finally gives the original field E of the wave when the rings expand to infinity.

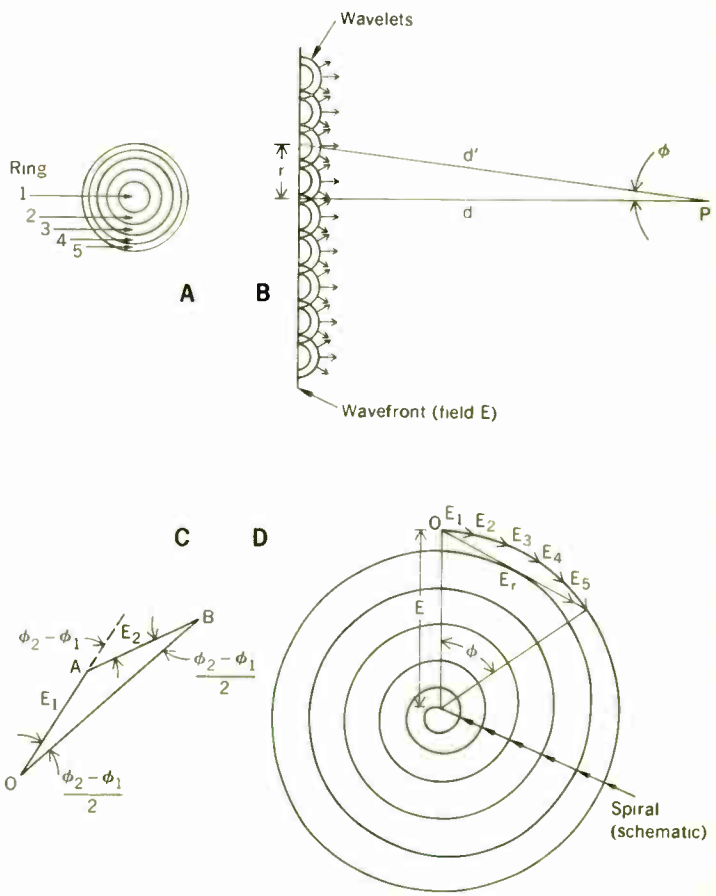
The phase of the field from the ring with radius r with respect to field E_1 from the field from the center element is

$$\phi = \frac{2\pi(d' - d)}{\lambda} \approx \frac{\pi r^2}{\lambda d}$$

Fresnel could have found the field $E_r = E_1 + E_2 + E_3 + E_4 + E_5$ due to all the rings with radii less than r. Figure 13(D) shows that it is

$$E_r = E\phi = E \frac{\pi r^2}{\lambda d} \text{ volts}$$

FIGURE 13. Propagation of light waves. A—Front view of wave. B—Side view of wave. C—Addition of two fields. D—Addition of wavelet fields.



or

$$\frac{E_r}{E} = \frac{\pi r^2}{\lambda d} *$$

This important formula will now be used to find the ratio of the power received by a light absorber with

* Since E_r (which is the graphical sum of E_1, E_2, E_3, E_4, and E_5) approaches $5E_1$ for very small values of ϕ, E_1 must be proportional to $1/\lambda$; hence the field of a Huygens source is proportional to $1/\lambda$, another very useful conclusion.

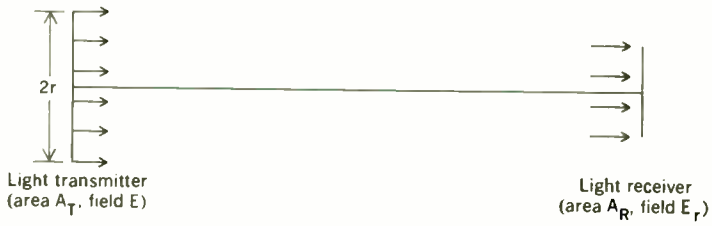
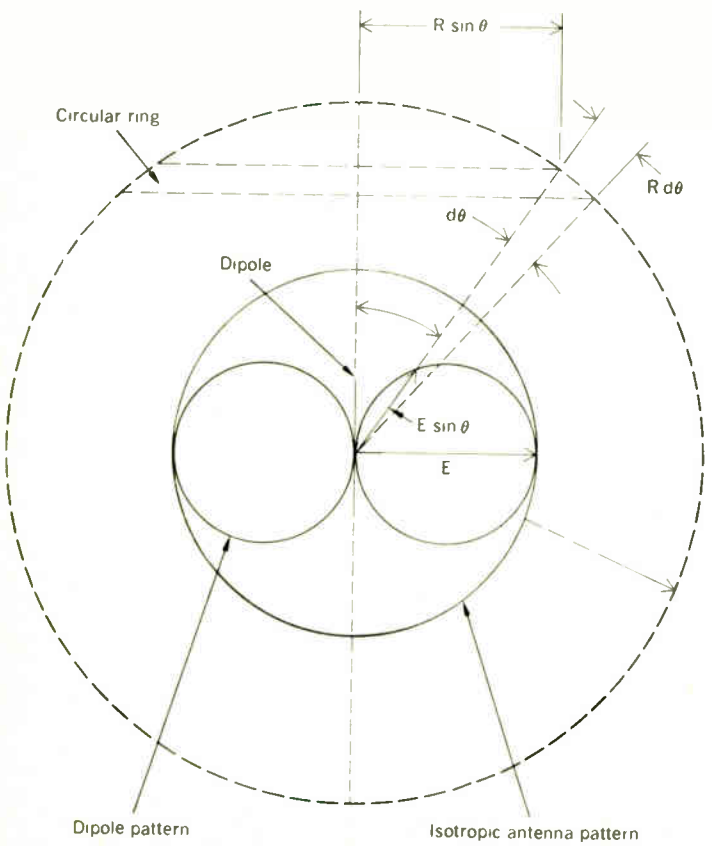


FIGURE 14. Propagation of light between a light transmitter and a light receiver.

FIGURE 15. Relation between the gain and the effective area of an antenna.

FIGURE 16. Gain of a small dipole.



area A_R and the power transmitted by a light transmitter with area $A_T = \pi r^2$. Figure 14 shows the light circuit. The received and transmitted power flows per unit area are proportional to the square of the light voltages. Hence,

$$\frac{P_R}{P_T} = \left(\frac{E_R}{E}\right)^2 \frac{A_R}{A_T} = \left(\frac{\pi r^2}{\lambda d}\right)^2 \frac{A_R}{A_T} = \frac{A_T A_R}{\lambda^2 d^2} \qquad (15)$$

Since light waves and radio waves obey the same laws, A_T and A_R may be considered to be the effective areas of radio transmitting and receiving antennas; this formula is identical to my transmission formula for a radio circuit[1] and it is interesting that it could have been derived long before Maxwell published his theory of electromagnetic waves in 1864 and Hertz discovered radio waves in 1888. Also, the formula applies to acoustic waves. Hence, there should be a derivation independent of electromagnetic theory.

Relation between the gain and effective area of an antenna. Figure 15 shows a hypothetical isotropic antenna that radiates a uniform field in all directions. The power flow at distance d is equal to the transmitted power P_T divided by $4\pi d^2$, the area of a surrounding sphere, and the received power is

$$P_R = P_T \times \frac{1}{4\pi d^2} \times A_R \text{ watts}$$

Replacing the isotropic antenna with a transmitting antenna of area A_T and using Eq. (15), we have, for the received power,

$$P_R' = P_T \frac{A_T A_R}{d^2 \lambda^2} \text{ watts}$$

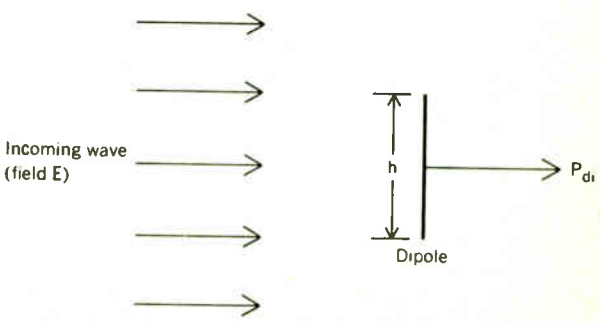
The gain g of the new transmitting antenna, defined as P_R'/P_R, is

$$g = 4\pi \frac{A_T}{\lambda^2} \qquad (16)$$

and this important formula applies to receiving antennas as well.

The gain of a small dipole. A wire carrying an alternating current radiates power. Hertz showed this experimentally years ago and later work showed that the radiation field in a direction is proportional to the projected length of the wire in that direction. Figure 16 shows the radiation patterns of a current element or small dipole (cross section of two circles with radii $E/2$) and an iso-

FIGURE 17. Effective area and radiation resistance of a small dipole.



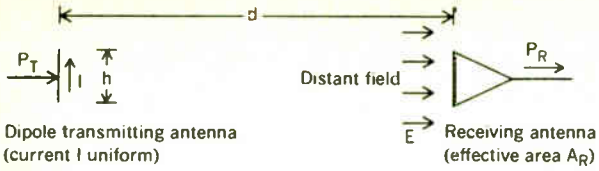


FIGURE 18. The distant field E from a small dipole with uniform current I.

tropic antenna (cross section of a circle with radius E). E is the field at a large distance R. The broken-line circle with radius R is the cross section of a sphere. The surface of the sphere is divided into circular rings with area $2\pi R \sin\theta \times R\,d\theta$. The power flow through the ring due to the dipole is

$$dP_{\text{dipole}} = \frac{\frac{1}{2}(E\sin\theta)^2}{120\pi} \times 2\pi R\sin\theta \times R\,d\theta$$

$$= \frac{\frac{1}{2}E^2}{120\pi} \times 2\pi R^2 \sin^3\theta\,d\theta$$

The power radiated by the dipole is then

$$P_{\text{dipole}} = \frac{\frac{1}{2}E^2}{120\pi} 2\pi R^2 \times \int_0^{\pi} \sin^3\theta\,d\theta = \frac{\frac{1}{2}E^2}{120\pi} 2\pi R^2 \times \frac{4}{3}$$

The power radiated by the isotropic antenna is

$$P_{\text{iso}} = \frac{\frac{1}{2}E^2}{120\pi} 4\pi R^2$$

The gain of the dipole is then

$$g_{\text{di}} = \frac{P_{\text{iso}}}{P_{\text{di}}} = 1.5 \qquad (17)$$

Effective area of a small dipole (Fig. 17). Equations (16) and (17) give

$$A_{\text{di}} = g_{\text{di}} \times \frac{\lambda^2}{4\pi} = 1.5 \times \frac{\lambda^2}{4\pi} \approx \frac{\lambda}{2} \times \frac{\lambda}{4} \text{ square meters} \qquad (18)$$

Note that this area is independent of the dipole length h (which is always small compared with the wavelength).

Radiation resistance of a small dipole. A small dipole with alternating current radiates power and, therefore, it has a resistance called the radiation resistance.

The power received by the dipole shown in Fig. 17 is

$$P_{\text{di}} = \frac{\frac{1}{2}E^2}{120\pi} A_{\text{di}} = \frac{\frac{1}{2}E^2}{120\pi} \times 1.5 \times \frac{\lambda^2}{4\pi} \text{ watts} \qquad (19)$$

The dipole is equivalent to a generator with resistance R_{ra} and induced voltages $E \times h$; the available power is

$$P_{\text{di}} = \frac{\frac{1}{2}(Eh)^2}{4R_{\text{ra}}} \text{ watts} \qquad (20)$$

Equations (19) and (20) give

$$R_{\text{ra}} = 80\pi^2 \left(\frac{h}{\lambda}\right)^2 \text{ ohms} \qquad (21)$$

Note that the radiation resistance is small when h/λ is small, and this makes it difficult to construct an efficient matching circuit to an output line.

The distant field from a small dipole with uniform current. For the circuit shown in Fig. 18, Eq. (15) gives

$$P_R = P_T \frac{A_{\text{di}} A_R}{d^2\lambda^2} \text{ watts} \qquad (22)$$

We also have

$$P_R = \frac{\frac{1}{2}E^2}{120\pi} A_R \text{ watts} \qquad (23)$$

and, since I is maximum value of the alternating current,

$$P_T = \frac{1}{2}R_a I^2 \text{ watts} \qquad (24)$$

Equations (18) and (21)–(24) give

$$E = 60\pi \frac{h}{d\lambda} I \text{ volts per meter}$$

Summary

Assuming unobstructed free-space transmission and using for the most part only mathematics taught in many high schools, it has been possible to derive the fundamental properties of radio antennas and radio transmission.

Dr. J. R. Pierce got me started on this article in the fall of 1969 by encouraging his co-workers to find a simple derivation of my old transmission formula. One of these co-workers, C. L. Ruthroff, mentioned the old work on light waves by Huygens and Fresnel and this set me off in the right direction. Dr. S. A. Schelkunoff suggested that I simplify the discussion by deleting the complex algebra in the original version. Lou Mitchel, physics teacher at Rumson High School, read the paper with one of his students and recommended its publication.

REFERENCE

1. Friis, H. T., "A note on a simple transmission formula," *Proc. IRE*, vol. 34, pp. 254–256, May 1946.

Harald Trap Friis (F) was born in Naestved, Denmark, in 1893. He received the electrical engineer and doctor of science degrees from the Royal Technical College, Copenhagen, in 1916 and 1938, respectively. In 1919 he moved to the United States and, after a period of study at Columbia University, he joined the Western Electric Company's Research Department, which was later to become Bell Telephone Laboratories, Inc. He was made director of radio research in 1945. In 1952 he became director of research in high frequency and electronics. During his career with the Bell System he contributed substantially to almost every aspect of the radio art, including vacuum tubes, the design of the first commercial superheterodyne receiver, noise, antennas and propagation, radar, and microwaves. Following his retirement from Bell Laboratories in 1958, he served until 1968 as consultant to the Hewlett-Packard Company. At present he is doing consulting work from his home at Rumson, N.J.

Dr. Friis is a member of the American Section, International Scientific Radio Union; and the Danish Academy of Technical Sciences. He was awarded the IRE's Morris Liebmann Memorial Prize (1939) and Medal of Honor (1955), the Franklin Institute's Stuart Ballintine Medal (1958), and the IEEE's Mervin J. Kelly Award (1964). He received the Danish decoration "Knight of the Order of Dannebrog," presented by King Frederick IX, in 1954, and the Valdemar Poulsen Gold Medal, presented by the Danish Academy of Technical Sciences, also in 1954.

Effective measurements using digital signal analysis

Digital hardware and the fast Fourier transform are not enough to guarantee successful signal processing. The user must also understand the nature of the measurement process itself

Peter R. Roth *Hewlett-Packard Company*

In order to be effective, measurements must be unambiguous and quantitative, which is only possible when such measurements yield numerical values that can attach to a given aspect of the system being measured in only one way and have only one interpretation. This article discusses the use of correlation, transfer, and coherence functions in obtaining such measurements.

Over the past few years, the significant reduction in the cost of both small computers and special-purpose digital hardware has resulted in a blossoming of digital signal processing as a useful measurement procedure. At the same time, the introduction of the fast Fourier transform (FFT) has made it practical to accomplish many signal-processing operations that formerly required an impractical amount of computation time. The combination of reduced hardware cost and the FFT has taken such concepts as correlation and power spectrum out of the textbook and made them into effective measurement tools.[1-3] The reality of this situation is reflected in the number of digital correlators and Fourier analyzers available today.

In order for the potential user to exploit fully this available measurement technology, it is important that he have a solid insight into the analysis process that he is to use. Such an understanding must go beyond simply a knowledge of the mathematics used, and its internal meaning, to a physical feel for the significance of the measurement process. When such an understanding exists, the user will be able to obtain quantitative and unambiguous results. Without this insight, there is a very real danger that the result of powerful digital signal analysis procedures such as correlation will yield qualitative and ambiguous results. I suspect that when such ambiguous results are obtained the measurement is often put off as being too "noisy" or too difficult for even these sophisticated analysis tools. This is unfortunate, for often these results are not due to deficiencies in these techniques, but in a failure fully to appreciate and exploit the theory available. Although the foregoing statements certainly apply to the art of autocorrelation and power-spectrum measurements, in this article we shall concentrate on the measurement of the cross-relationship between two signals.

When a measurement of the relation between two signals (to define either some similarity or the characteristic of the transmission path between them) runs into certain difficulties, correlation or related techniques may suggest themselves as solutions. This most often happens when the signals to be measured are ill-conditioned. Either there is noise masking the signals of interest or the signals themselves are noise-like and will not yield to conventional magnitude-, phase-, or time-measurement procedures. For example, measuring the relative time delay of a complex pulse transmitted over a noisy channel will not yield to techniques using electronic counters. When situations such as this develop, correlation usually is the first technique tried.

To use correlation effectively or to choose some extension of correlation to solve the measurement problem, it is important to understand precisely what correlation is. To most engineers correlation suggests two meanings. There is the plain-English meaning of "similarity between two things," which is too imprecise to be of real aid in a technical sense. On the other hand, the mathematical definition is too abstract to help develop the measurement sense we desire. What we shall attempt to do here is bridge the gap between the plain-English meaning of correlation and its mathematical statement as the correlation function, and show how the connotation of the word may be in conflict with its mathematical denotation. We intend not only to resolve this conflict, but to clarify the significance of certain measurement tools related to the correlation function.

It turns out that there are really two distinct quantities that one usually desires to measure when the correlation function is applied to a problem. Consider the generalized measurement situation shown in Fig. 1, where there are two accessible measurement points X and Y. X and Y are related by some linear transfer quantity H, and Y is perturbed by some uncorrelated noise source N. As we shall see, most two-point measurements can be modeled in this way. Even situations where H is not a physical transmission path can be structured in the manner of Fig. 1 as long as it defines a linear dependence of output on input. The two basic measurements that can be made on this model involve the transmission path H and the degree to which the uncorrelated noise N affects the output Y. The first is a measurement of the nature of

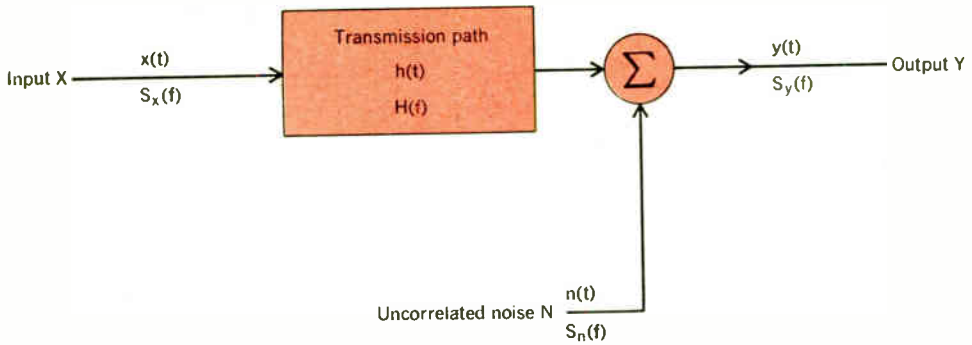


FIGURE 1. Diagram of the generalized measurement problem to be solved using techniques related to correlation. Most measurement problems involving two signals or a stimulus and a response can be modeled in this way.

the relationship of X and Y; the second is a measurement of the degree of causality between X and Y.

The introduction of the words causality and relationship might at first appear further to confuse the meaning being sought. Their introduction, however, gives voice to certain concepts that are not clear in the single word correlation. For example, the communication engineer measuring a transmission line is interested in the nature of the transmission path in the sense of what the transmission characteristic does to the signal. A neurobiologist, on the other hand, may apply a stimulus to a subject and desire to know only if there is a response and its strength. He might very well not desire to know the way in which the stimulus is transmitted, but only the response strength against the background of noise. These are measurements of relationship and causality respectively. Both of these concepts are implied in the word "correlation." Hence, it is extremely important in the proper execution and interpretation of correlation measurements for the user to realize that there is a significant difference between a measurement of relationship and a measurement of causality.

Let us start by considering two functions that can be used to measure the relation between two signals. The first is an estimate of the correlation between X and Y over a period of time T given by

$$R_{yx}(\tau) = \int_0^T y(t)x(t + \tau)\, dt \qquad (1)$$

where R_{yx} is simply the averaged product of y lagged with respect to x. When the value of the cross correlation is high for some value of the lag τ, it can be said that x and y are similar, in some sense, at this lag value. But exactly what the significance of a given value of $R_{yx}(\tau)$ is cannot easily be determined. It is frequently assumed that by normalizing the cross-correlation function the problem of interpretation can be solved. The cross correlation is normalized by dividing by the geometric mean of the input and output mean-square value, giving

$$\rho_{yx}(\tau) = \frac{R_{yx}(\tau)}{\sqrt{R_{xx}(0)R_{yy}(0)}} \qquad -1 \leq \rho_{yx}(\tau) \leq 1 \qquad (2)$$

where ρ_{yx} is the normalized cross correlation and $R_{xx}(0)$ and $R_{yy}(0)$ are the autocorrelation functions at X and Y for zero lag.

In reality, normalization only removes the problem of how to interpret the cross-correlation function for two values of ρ_{yx}. When the cross correlation is ± 1.0, the interpretation is clear. For that value of lag, the transmission path has no loss and there is no contaminating noise. When ρ_{yx} is zero, there is effectively no transmission for this lag because either the system gain is 0 or there is an infinite amount of contaminating noise. However, when ρ_{yx} is less than 1.0 but greater than 0, the significance of the result is less clear.

Does such a value mean that the transmission path has a gain less than one for this lag or does it imply that some contaminating noise has reduced the degree of causality between input and output? The answer is not clear. Thus, simply normalizing the correlation function does not change a qualitative measurement to a quantitative measurement that differentiates between the relationship and the degree of causality between X and Y.

To resolve this problem, it is necessary to introduce some useful frequency-domain functions. The cross power spectrum, which is the Fourier transform of the cross-correlation function, is the key element in this approach:

$$G_{yx} = F\{R_{yx}\} \qquad (3)$$

However, more clarity in development can be obtained if G_{yx} is developed from its component spectrums. For the variables of Fig. 1, two linear spectrums can be found that are the Fourier transforms of $x(t)$ and $y(t)$:

$$S_x = F\{x(t)\} \qquad (4)$$

$$S_y = F\{y(t)\} \qquad (5)$$

From these linear spectrums, three power spectrums may be computed. These are the input power spectrum,

$$G_{xx} = S_x S_x^* \qquad (6)$$

the output power spectrum,

$$G_{yy} = S_y S_y^* \qquad (7)$$

and the cross power spectrum,

$$G_{yx} = S_y S_x^* \qquad (8)$$

where the asterisk in these equations indicates a complex conjugate. The output power spectrum can be broken down into its two component parts, one due to the input and one due to the uncorrelated noise source:

$$S_y = HS_x + S_n \tag{9}$$

Substituting (9) into (8), the cross power spectrum can be written as

$$G_{yx} = HG_{xx} + G_{nx} \tag{10}$$

Inversely transforming Eq. (10) yields

$$R_{yx} = h \star R_{xx} + R_{nx} \tag{11}$$

for the cross-correlation function, where the star indicates the convolution of the two time functions.

Equations (10) and (11) greatly clarify the meaning of cross correlation and the measurement-interpretation problems associated with it. The first term in each of these relations represents a measurement of the transmission path, and the second term is related to the degree to which the uncorrelated noise from N reduces the causality between X and Y. The two equations reveal that either a cross-spectrum or a cross-correlation function computed from the same data are exactly equivalent and that neither is statistically better than the other. These equations also reveal that a single cross measurement is made up of two elements—the first due to the system transfer characteristic, and the second due to noise uncorrelated with the input.

In many cases, either a random input signal or an uncorrelated noise term will require ensemble averaging or signal integration of some form to smooth the measurement result. It is often falsely assumed that the integration in the cross-correlation function provides some major advantage over the cross spectrum in terms of statistical stability. This is not the case. To provide statistical stability and lower the variance in a result, it is necessary to divide the data into small sections from which estimates of either R_{yx} or G_{yx} are computed. If these estimates are then averaged, the result (denoted by $\overline{R_{yx}}$ and $\overline{G_{yx}}$) provides a more reliable measurement. From the same amount of data, either cross measurement provides the same degree of certainty in the result.

It is interesting to observe the effect of a large number of ensemble averages on the measurement of a simple cross term. The longer the average, the more nearly the cross term $\overline{R_{nx}}$ or $\overline{G_{nx}}$ approaches zero since n and x are assumed to be uncorrelated. While the measurement of the cross term becomes more certain, the degree to which the noise contaminates the measurement becomes less visible. At the same time, it becomes more difficult to answer the question, "To what degree is y correlated with x in the sense of a causal relationship and to what extent does the noise contaminate y?" What a well-averaged cross measurement does approach is an approximation to the system transfer relation. However, it is only an approximation for two reasons: first, because it contains a term dependent on the input (R_{xx} or G_{xx}); second, because there is no way to determine the residual effect of the noise term (R_{nx} or G_{nx}). In the following three sections, we examine some solutions to the problem of identifying a system and proving the assumption of causality in a measurement by extending the concept of correlation to transfer- and coherence-function measurements.[4,5]

The transfer function

First, let us consider the case in which a measurement of the transfer relation of a system, as modeled in Fig. 1, is desired, and in which the nature of the stimulus or environment requires that some form of correlation be used. Assume either that S_n is small or $\overline{G_{yx}}$ has been smoothed enough to make $\overline{G_{nx}}$ negligible (a test for this condition is considered in the next section). The cross-spectrum and cross-correlation terms are then given by

$$\overline{G_{yx}} = H\overline{G_{xx}} \tag{12}$$

$$\overline{R_{yx}} = h \star \overline{R_{xx}} \tag{13}$$

The assumption usually made in this case is that X is white random noise. If this assumption is valid, then G_{xx} is a constant, R_{xx} is a delta function, and the cross measurements in (12) and (13) differ from the true transfer relations by only a multiplicative constant. Typically, in practice G_{xx} is not a constant, and two types of errors can appear in these measurements. The first is deterministic error due to the nonwhiteness of G_{xx}, and the second is a variance due to any randomness in G_{xx} or residual G_{nx}.

The deterministic error due to G_{xx} not being flat is easy to understand. However, its effect is frequently underestimated. Noise generators with truly flat spectrums are hard to find and the natural stimulus in most measurement situations is usually far from flat. Even when a precision noise source is used as a stimulus in an experiment, spectrum shape is frequently lost when the generator is coupled into the experiment.

The variance error due to G_{xx}, on the other hand, is not so obvious. If the input to the system under test is random, then the measurement of the cross spectrum or correlation will have a variance due to the input's random nature. Thus, our measurement of the transfer relation will have variability in spite of the fact that H, the quantity we are trying to measure, is nonrandom. To reduce this error, some sort of smoothing must be performed on the measurement, either by ensemble averaging of estimates of G_{yx} or R_{yx}, or by using a smoothing function on one estimate alone.

Both the deterministic and variance errors can be nearly eliminated if the transfer function is first computed by dividing the input power spectrum into the cross power spectrum. This gives the transfer function directly:

$$H = \frac{\overline{G_{yx}}}{\overline{G_{xx}}} \tag{14}$$

The impulse response of the system is easily found by inverse-transforming (14) to obtain

$$h = F^{-1}\left\{\frac{\overline{G_{yx}}}{\overline{G_{xx}}}\right\} \tag{15}$$

It is clear that normalizing by G_{xx} removes the deterministic error in the measurement of the transfer relation by taking into account the actual shape of the drive spectrum G_{xx}. What is not so clear is that the first-order effects of randomness are also removed. If G_{xx} and G_{yx} are measured from the same sample records and x and y are related by a linear system, then x and y have essentially the same random variation. Thus in Eq. (14) the randomness is canceled out and the measurement of the transfer relation is indeed nonrandom. The only precaution that must be taken is that G_{xx} and G_{yx} are from simultaneously sampled records. If not, the random variation will not be the same in each quantity and the

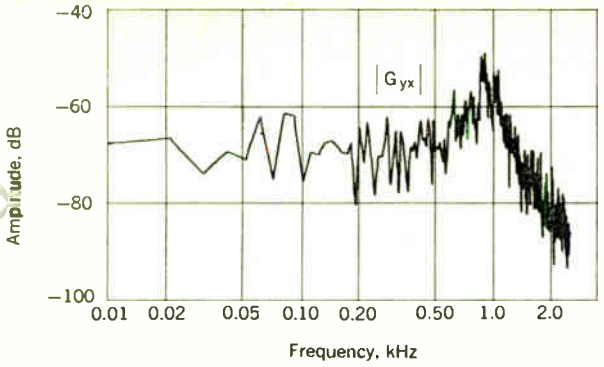


FIGURE 2. The cross power spectrum for an underdamped second-order system with white random noise excitation. The measurement, as with all others in this article, was made with a Hewlett-Packard 5450A Fourier analyzer, which implements correlation and power-spectrum measurements using an HP computer and a set of routines built around an FFT algorithm.

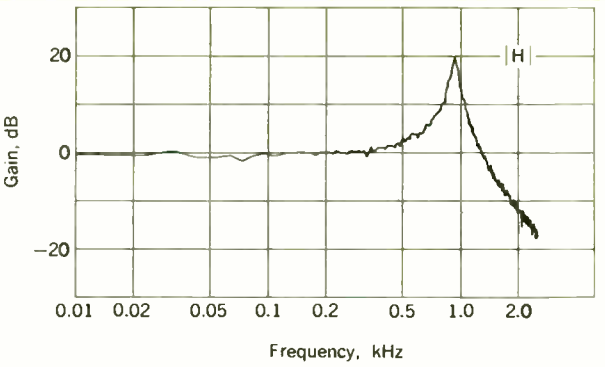
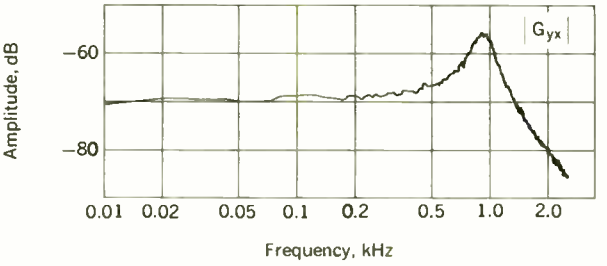


FIGURE 3. Transfer function computed from the same data as in Fig. 2. Variability error is greatly reduced, and the ratio of low-frequency response to peak response is 20 dB, indicating a Q of 10 for the system.

FIGURE 4. The cross power spectrum for the system of Fig. 2 from 100 times the data (40.96 seconds). Variability is about the same as the transfer function computed from 0.4096 second of data, and the Q is only 5.6 (computed from a low-frequency-to-peak ratio of 15 dB). This error is attributed to the nonwhite character of the input caused by the reactive nature of the noise-generator load.



cancellation will not occur. The most important result of Eqs. (14) and (15) is that, to achieve a given level of accuracy, far fewer estimates are needed; or, with a limited amount of data, as in a nonstationary process, the result may be obtained with greater accuracy.

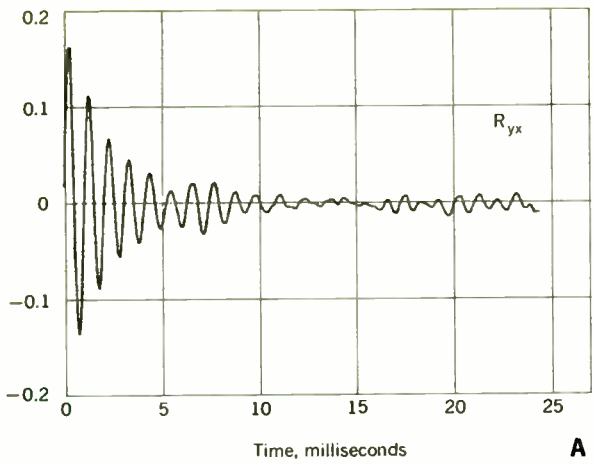
Let us demonstrate the significance of the foregoing statements by examining the result of some measurements made on a second-order, underdamped system using random noise as a stimulus. The random source used had a 3-dB cutoff of 5 kHz and a 40-dB-per-decade roll-off. In Fig. 2, the cross spectrum resulting from the average of four estimates of G_{yx} is shown. Each spectral estimate was computed on the basis of 102.4 ms of data, so the basic resolution of measurement was about 10 Hz and the total record length was 409.6 ms. In Fig. 3, the same 409.6 ms of data is shown with the transfer function H computed as in Eq. (14). The cancellation of statistical variation is clearly demonstrated in Fig. 3, where the variation in the measurement is less than ± 1 dB compared with ± 8 dB in Fig. 2. The point is also made in Fig. 4, where a cross spectrum is computed from 400 spectral estimates (40.96 seconds of data). Here the variation for 100 times the data is about the same as in the transfer-function measurement of Fig. 3. The deterministic error is quite clear in Fig. 4, in spite of the fact that the source was supposed to be flat and wide-band. (The 600-ohm source impedance of the generator made the coupling network frequency-dependent.)

Note that the peak of the cross spectrum in Fig. 4 is 14 dB above the low-frequency value. This compares to the 20-dB peak value of Fig. 3, which is in perfect agreement with the actual Q value of 10 for the network. A similar comparison can be made for the cross correlation computed from G_{yx} and the impulse response computed from Eq. (15). Figure 5(A) shows R_{yx} computed from 400 ms of data and Fig. 5(B), the impulse response h from the same data. It is sufficient to note that the envelope of h is a monotonically decaying function and that the envelope of R_{yx} is not.

The question of the validity of the assumption that G_{nx} is small in the measurements just described has been ignored. It is important to consider that, as the level of the noise added by the uncorrelated source N grows, the variance in the measurement increases. In this situation, the advantages of transfer-function measurement over the cross-spectrum approximation are not as great as in the example given. When the signal-to-noise ratio is low, a number of averages must be taken to reduce the variance due to N and the cancellation of the input randomness is less significant. It is clear then that it is important to be able to determine the degree to which N affects the system output. In the next section, we shall examine the question of determining causality (i.e., the amount of contaminating noise) in a measurement.

The coherence function

Often the measurement that is desired is not of the system transfer characteristic, but rather of the degree to which the system output is the product of the input versus some other uncorrelated source. When we speak of how well the input and output are correlated, it is frequently in the context of input–output causality. In this sense, consider measuring the "correlation" between input and output using Eqs. (10) and (11). What we are asking is, "What percent of the cross spectrum or correlation is due to the noise term?" But it is impossible to separate the system-related term HG_{xx} from the noise term, G_{nx}. In fact, if we are able to average the result well enough, the noise term will ideally go to zero.



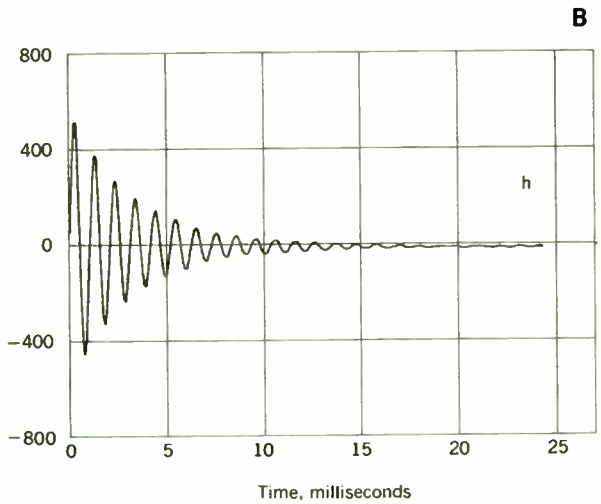


FIGURE 5. A.—The cross-correlation function from the data of Fig. 2. The nonmonotonic envelope and roughness result from deterministic and random error. B—The impulse response from the data of Fig. 3; note the clearly defined monotonic envelope.

Thus the result derived from a cross-correlation function or cross-spectrum measurement will describe the effect of the system on the input, independent of contaminating noise, and will not give a measure of how well the input and output are "correlated," in the plain-English sense of the word. If the functional form of the input and transfer relation are known then, in principle, it would be possible to apply statistical tests to a measurement of G_{yx} or R_{yx} to determine the effect of the noise term. But, in practice, such a procedure is cumbersome. Fortunately, a combination of the cross measurement with two auto measurements, called the coherence function, resolves this problem.[6–8]

The coherence function is defined by

$$\gamma^2 = \frac{|\overline{G_{yx}}|^2}{\overline{G_{xx}}\,\overline{G_{yy}}} \tag{16}$$

To interpret this ratio of averaged spectrum, let us examine some of its component parts. If the output power spectrum is written using Eqs. (7) and (9), we obtain

$$G_{yy} = H^*HS_xS_x^* + H^*S_x^*S_n + HS_xS_n^* + S_nS_n^* \tag{17}$$

Assuming, as before, that the sources X and N are not correlated, upon averaging we obtain

$$\overline{G_{yy}} = |H|^2\overline{G_{xx}} + \overline{G_{nn}} \tag{18}$$

The first term of the right-hand expression in (18) represents the power at the system output resulting from the input, while the second term is the noise power. Using the same assumption that the uncorrelated cross term is zero, from Eq. (10) the magnitude squared of the measured cross power spectrum $\overline{G_{yx}}\,\overline{G_{yx}}^*$ is

$$|\overline{G_{yx}}|^2 = |H|^2\overline{G_{xx}}^2 \tag{19}$$

Thus Eq. (16) can be rewritten as

$$\gamma^2 = \frac{|H|^2\overline{G_{xx}}}{|H|^2\overline{G_{xx}} + \overline{G_{nn}}} \tag{20}$$

Using Eq. (20), a simple interpretation of the coherence function is possible. At a given frequency, γ^2 is the fraction of the power at the system output that is due to the input. If $\overline{G_{nn}}$ is zero, γ^2 is 1, indicating a perfect linear, nonnoise-contaminated relation between input and output. Yet, if $|H|^2G_{xx}$ (the output term due to the input) is small compared with the noise, γ^2 will tend to zero. *γ^2 is a value lying between 0 and 1.0 that gives an unambiguous indication of the causal relation between input and output.*

The coherence might be misinterpreted as simply the frequency-domain analog of the normalized correlation function. This is certainly not true. The difference lies in the fact that the correlation function is normalized only to the mean of the input and output power whereas the coherence function is normalized at each frequency independently. Thus it is possible for the coherence function to represent the causality between input and output without reference to system transfer function, although the functional shape of the normalized cross correlation or spectrum will represent the system transfer function.

The normalization of the coherence function at each frequency allows it to be used to find some interesting power spectrums. Hence, $\overline{G_{yy}}(1 - \gamma^2)$ is the power spectrum at the output due to the system noise and $\overline{G_{yy}}\gamma^2$ is the signal power at the system output. The signal-to-noise ratio as a function of frequency is then

$$\frac{S(f)}{N(f)} = \frac{\gamma^2}{1 - \gamma^2} \tag{21}$$

The fraction expressed in (21) is even a more useful definition of the causality between input and output than γ^2. It is also important to note that γ^2 gives an important check on the validity of the measurement of a transfer function. If γ^2 is zero there is no input–output relation and any measurement of H is invalid. Bendat[9] covers this topic in some detail. The importance of γ^2 and Eq. (21) arises from the fact that, although the SNR cannot be directly measured, γ^2 is simply computed from some easily measured spectrums.

We can illustrate the significance of the coherence function with an extension of the measurement used for the transfer function example. In Fig. 6, the cross-correlation function is plotted for a basically similar set of conditions as in Fig. 5, except that the signal from N is uncorrelated random noise with a 3-dB bandwidth of 500 Hz. As was previously predicted, the shape of the correlation function looks much the same in both

Fig. 5(A) and Fig. 6. It is, in fact, impossible to decide from Fig. 6 alone to what degree the system noise contaminated the signal. Figure 7 shows the coherence function for this situation, and we can see that between zero and 500 Hz the output of the system is strongly contaminated by noise and that above 500 Hz the input rather than the noise contributes to most of the output.

The plot of the SNR for this system is shown in Fig. 8. Here again we observe that below 500 Hz the system noise predominates, as expected from the nature of the experiment. Figures 6 through 8 demonstrate that in most measurement situations the cross-correlation function does not measure the "correlation" in a quantitative way, in the sense of causality between two signals, and that one must use the coherence function to obtain this sort of measurement. The same statements can be made with respect to the cross spectrum.

The coherence function, like the transfer function, is basically defined for a linear, stationary system. However, if the system is nonlinear, the products of the nonlinearity are treated as system noise and reduce the value of coherence. If the system is assumed to be nonstationary, neither correlation nor coherence will yield a completely correct measurement. However, the coherence yields a quantitative estimate of the correlation with fewer estimates than required for the correlation function because the randomness is normalized out. Thus, if γ^2 is 1 it may be clearly said that input and output are related by a noise-free linear system whose gain is defined by the transfer function. On the other hand, if γ^2 is zero the input and output of the system are not related at all. As pointed out by Bendat,[9] if γ^2 is between 0 and 1, three conditions can exist: (1) Extraneous noise is present in the measurement. (2) The system relating X and Y is nonlinear. (3) Y is an output resulting from an input X as well as other inputs.

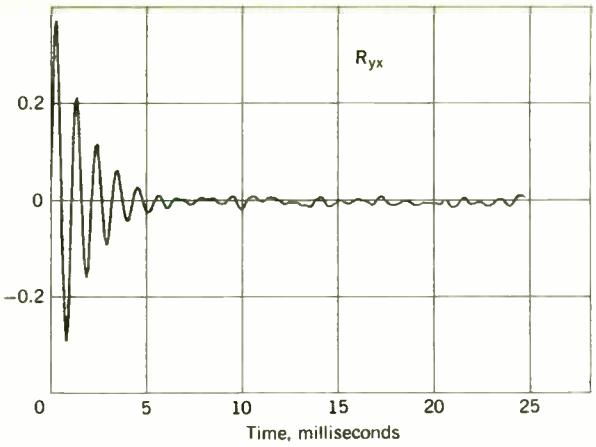
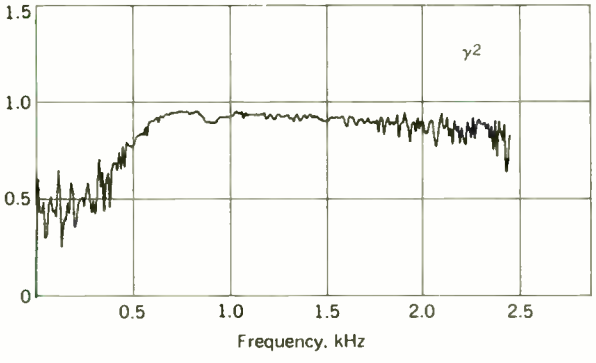


FIGURE 6. Cross correlation for the system measured in Fig. 2 with uncorrelated noise added as in Fig. 1. Sample rate and sample record length are the same as in Figs. 2 through 5, except that 4.56 seconds of data were used. A comparison with Fig. 5(A), where there is no system noise, shows that it is impossible to tell from this plot if there is a lack of causality between the input and output. The difficulty in using these data to determine the degree of causality lies in the fact that R_{yx} mainly reflects the transfer relation between input and output.

FIGURE 7. The coherence function for the system of Fig. 6. These data clearly show that, from direct current to 500 Hz, there is a lack of correlation between input and output. At low frequencies, the coherence is 0.5, which indicates that 50 percent of the output energy comes from the input. Coherence is extremely useful, since the functional shape of the transfer relation is normalized out and only the coherence between input and output remains.



Path-delay measurements

In the previous two sections, we have described procedures for making measurements of systems with physical inputs and outputs. It is important to realize, however, that the concepts of causality and relationship are more general than the applications to physical systems alone would suggest. It is not at all necessary for a measurement to involve a system with a physical input and output to apply the concepts of transfer and coherence functions. To illustrate this point, let us consider a measurement of differential path delay. This problem is generally considered to be a classic application for correlation. The usual situation for this type of measurement involves the comparison of two complex time records to determine, by cross correlation, the time delay between them.

Although the exact way in which a differential path delay may be generated can assume several forms, a typical case is depicted in Fig. 9. Here sources Y and X transmit similar pulses that are delayed from one another by some constant amount. In addition, the difference in transmission-path length introduces more relative delay. One further complication results from the fact that source Y takes two paths to the receiver. This situation is typical of the problems encountered in the loran navigation system, where both the ground wave and sky wave from one transmitter may be received, whereas only the ground wave from the second transmitter will be detected.*

In this situation, the benefit to be derived from correlation is that it should be possible to measure the time delay between the signals in spite of their complex form, a difficulty that cannot be handled by conventional techniques. To apply correlation to such a situation, some implementation of Eq. (1) is generally used. It may be applied in analog, digital, or hybrid form, or by

* The loran navigation system uses pulses transmitted with a fixed delay from two known locations. A receiver at an unknown location measures the departure time of the signal from the known delay and computes a hyperbolic line of position. Two such measurements yield the receiver position. This system is often confused by the arrival of two pulses from a single transmitter. One follows a direct path over the ground and the second a slightly longer one reflected off the ionosphere.

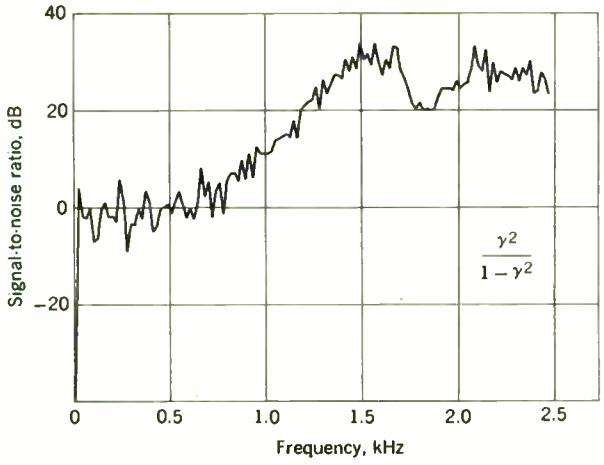


FIGURE 8. The signal-to-noise ratio computed from the data of Fig. 7.

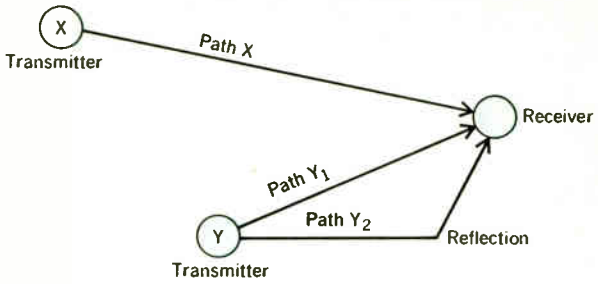


FIGURE 9. One form of the path-delay measurement problem. Transmitters X and Y radiate a pulse with a known delay, and by measuring the difference in transmission time the position of the receiver may be determined. Two possible delays are possible: Y_1X and Y_2X. Y_1 is the first pulse received, and Y_1X is the longest of the two delays. Since Y_1 and X are both unreflected paths, the longer Y_1X delay is the correct result for this experiment, with Y_2X representing a spurious result.

some analogous technique such as phase detection. The exact form of the implementation is unimportant. What is important is the interpretation of Eq. (1) relative to the measurement of the differential path delay. The usual reasoning is that, as the two signals are lagged in relation to each other, the average of their cross products, given by the cross-correlation function of Eq. (1), will approach a maximum when the two signals are best aligned. The value of lag at which the signals line up the best is then taken as the most probable value of delay for the system.

Such an interpretation is an oversimplification that may lead to significant errors. To understand why, we must again use the equation for cross correlation derived from the frequency-domain relations. Neglecting noncorrelated noise, which is an additive term, we have

$$R_{yx} = h \star R_{xx} \tag{22}$$

It is important to make explicit that although the concept of impulse response, represented by h in (22), is normally applied to a system that transmits energy, in this case we are using it to describe the *relationship* of the signals received from X and Y. In Eq. (22), h defines the relation between the signals at points X and Y and in no way applies to the transmission paths. What the correlation function R_{yx} describes is the relation between X and Y represented by h convolved with the signal autocorrelation function R_{xx}. The measurement of interest represented by h is confused by the irrelevant information contained in R_{xx}. Consider the situation in which both the signals and the transmission paths for X and Y in Fig. 9 are identical except that there is a perfect delay between X and Y. The relationship we desire to measure is an ideal delay τ_0, represented by a delta function at the delay $\delta(\tau_0)$. What cross correlation yields is the delay smeared by being convolved with R_{xx}. The solution to this problem is to calculate h rather than R_{yx} by first computing H from the ratio of G_{yx}/G_{xx} and then applying the inverse transformation as in Eq. (15).

The difficulties described in the measurement of differential path delay are exactly analogous to the deterministic and random errors associated with the measurement of the transfer function. Exactly the same benefits are obtained from measuring h rather than R_{yx} as are obtained from measuring H in the case of the transfer function. To see how much advantage the elimination of the deterministic error caused by R_{xx} can be, let's look at an actual measurement. The conditions of Fig. 9 were created by a double-pulse generator driving a resonant circuit to derive the double-path signal for Y. The delayed signal for X was created by driving a similar resonant circuit with a delayed pulse.

In this case, there are two possible delays that can be measured. One is between Y_1 and X and the second is between Y_2 and X. Since the Y_1 pulse arrives first, the delay Y_1X will be the longest and, as it is the delay for the unreflected path, it is the measurement that is sought. On the other hand, the delay Y_2X will be the shortest, and represents an incorrect result. Figure 10 shows the signals received from Y and X. The fact that Y is made up of a double pulse is clear when it is compared with X. In addition, the complexity of these signals suggests that R_{xx} will be complex and will hinder the measurement desired. The cross correlation R_{yx} is plotted in Fig. 11(A). Here we see one major peak representing one of the two possible delays. The complexity introduced by R_{xx} makes it impossible to tell which of the delays we are actually measuring. It is only when the measurement of h in Fig. 11(B) (derived from the inverse transform of G_{yx}/G_{xx}) is examined that we can observe the two delays. We now find that the peak revealed by cross correlation actually occurs for the incorrect short delay.

If one is to use cross correlation to measure differential path delay, then it is important to consider the true nature of the signals in the two channels. If there is only one delay possible, cross correlation will yield a correct answer, but the result will be at least as complex as the autocorrelation of one of the signals and may mask the true nature of the process being measured. If there is more than one delay possible (because of multiple paths), cross correlation may yield an incorrect result, as seen in the example just given. Although cross correlation can be used if enough caution is exercised in processing the results of an experiment, all ambiguity in the measurement can be eliminated if the impulse response is computed directly from the transfer function.

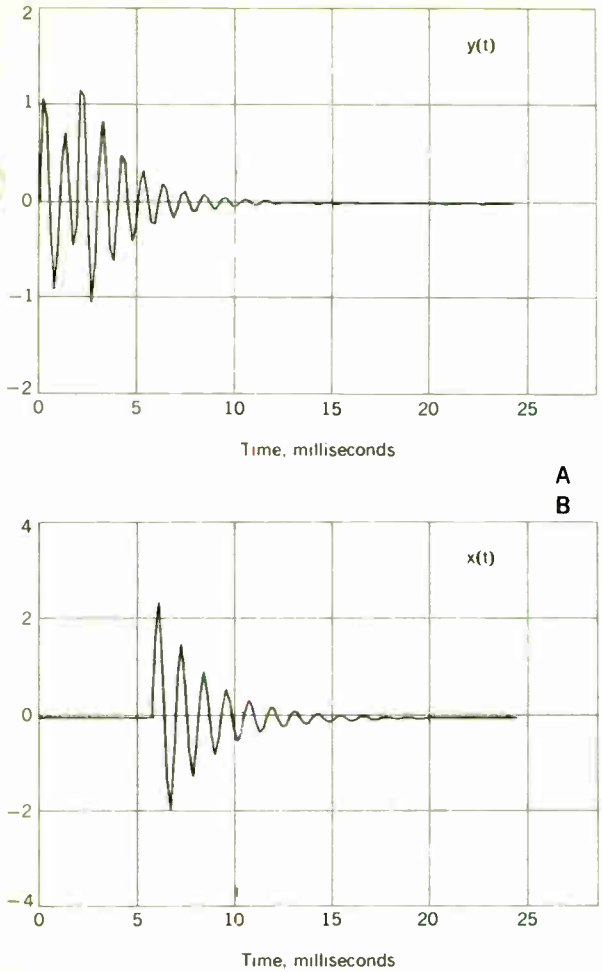


FIGURE 10. A—The double pulse received from transmitter Y for the situation shown in Fig. 9. B—The single pulse received from transmitter X.

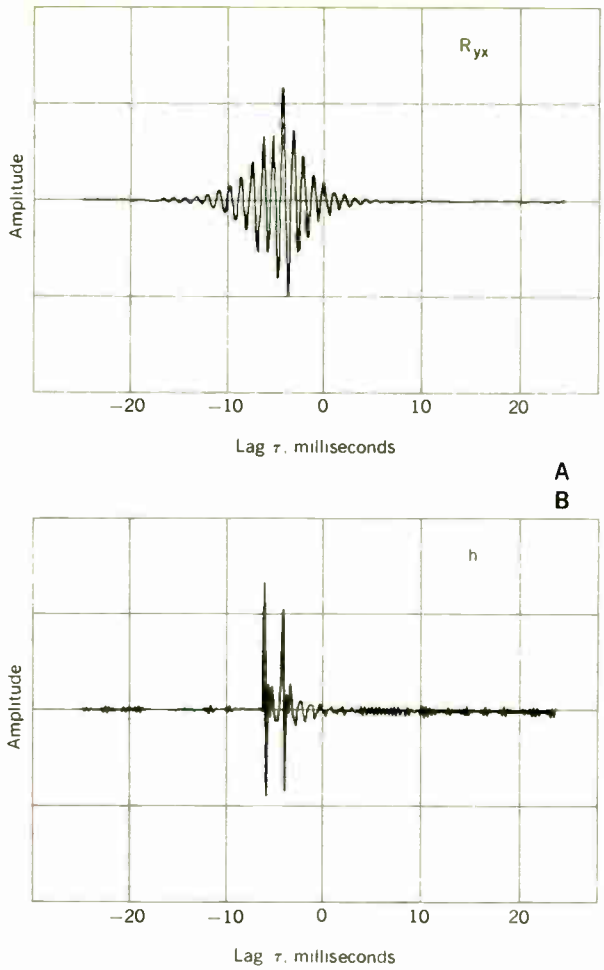


FIGURE 11. A—Cross correlation of the signals in Fig. 10. Note that, although two delays are physically present, only one (at −4.3 ms) is clearly resolved. B—The impulse response of the same signals as in A. Note that both delays are shown and that the correct longer delay (at −6.2 ms), not apparent in A, is now clearly shown.

Implementation

At this point, we have hopefully made the reader aware of the fact that it is possible to describe the meaning of correlation more precisely than "a measure of the similarity between signals" without becoming bogged down in mathematical abstraction. Although the differences in the techniques related to correlation that we have described are not complex, they can be subtle and deserve some final comment. The best way to clarify further the objective of this article is probably by tracing the line of reasoning that is needed to select the proper measurement procedure and hardware for a given problem.

The first element in such a line of reasoning should be to ask, "What is it that is to be measured?" This is not the obvious question that it appears to be, for it implies a differentiation between causality and relationship. For example, if the measurement to be made is in the nature of a system transfer function or path delay, then the measurement is one of relationship and one set of hardware performance characteristics must be considered. If, on the other hand, the measure we desire is the degree to which the input is contained in the output, a measure of causality in the sense of coherence is desired. For this result, a different set of hardware performance characteristics must be considered. Hence, the answer to the question of what is to be measured is one that has significant monetary consequence. How this question is resolved will determine whether the instrument to be used is a low-cost correlator, a cross-spectrum analyzer, or a more costly digital processor capable of performing transfer and coherence functions.

Consider the case in which the experimenter desires to measure the part of the output that is caused by the input. An example of this situation is the neurobiologist who is interested in the degree of response to a given stimulus. In this case, a measure of the transfer function of the system under test is not desired. What is wanted is the "correlation" between input and output in terms of signal-to-noise ratio. This is what we have chosen to call causality. Although the plain-English meaning of correlation might suggest that the correlation function is a definitive measure of the agreement between signals, we have shown that it is really a qualitative measure of the impulse response, and that coherence is the quantitative gauge actually desired. Equation (11) defines the cross correlation to be the convolution of the system impulse response with the autocorrelation of the input plus a noise-related term R_{nx}. As R_{nx} averages to

zero, the more statistically certain our measure of R_{yx} becomes, and the more we learn about the input R_{xx} and the system h. But at the same time, we become less certain of how much the response is contaminated by noise. Some improvement in this situation can be made by normalizing the cross-correlation function as in Eq. (2). In this case, the noise contained in the output power measure $R_{yy}(0)$ will change the scale of the normalized cross-correlation or cross-spectrum measurement and will allow a qualitative comparison of two measurements of an identical system. The major limitation here is that the functional form of the cross-correlation function is a measure of the input and the system transfer relation, and has nothing to do with the causality or signal-to-noise ratio. If a system or an experiment is modified to improve the degree to which the output rejects noise, the transfer function will change and it may be very difficult to interpret the change as reflected in the cross correlation.

The coherence function, on the other hand, is independent of both the system input and transfer function. Thus it measures the degree of noise contained in the system output in a quantitative manner. Using the coherence function, it is possible to say that one set of circumstances results in an output that contains more input than another. A system or experiment may then be optimized for the greatest response at the output due to the input in a direct and clearly interpretable way. If this is the reason for making the measurement, the improvement in interpretability that results from using the coherence function may well demand the extra cost that its implementation may involve.

A different set of solutions to the measurement problem arises when the quantity to be measured is that of relationship. In this situation, we must be very aware of the fact that the cross measurement R_{yx} or G_{yx} is a combination of the input and the system transfer relation. The result given by a single cross measurement in this case is more meaningful than in the case of coherence, and contains the quantity of interest, but is smeared or distorted by the nature of the input R_{xx} or G_{xx}. The degree to which the input function modifies the transfer relation will determine the accuracy given by the cross-correlation function. If the resulting accuracy is too low the transfer-function technique will have to be implemented at increased cost.

The accuracy of a measure of the transfer relation may be decreased by the nature of the input in two ways. As was pointed out in the section on transfer functions, even accurate noise generators will exhibit a departure from a flat spectrum and a delta-function autocorrelation function. One should note that such a situation makes the correlation function a qualitative measure of the system that depends on the level of the input and, to some degree, on the form of the input. Any measure of the transfer function or impulse response using the spectral ratio of Eqs. (14) and (15) results in a quantitative measure of correct scale and undistorted form. Even when the relationship sought is that of the path delay, the distortion can be profound. If the system has multiple delays or reflections and the duration of the input (the measured signal) autocorrelation is not shorter than the time lag between delays, shifts or incorrect measurements of the time lag can result.

Before deciding between the use of either the correlation or transfer function as the measurement technique, an evaluation of the degree of error that is introduced by the input autocorrelation function is needed. Here again, the data-processing cost compared with the accuracy and reliability of the result must be evaluated. The accuracy of a transfer relation may also be distorted by the variance added to the estimate of the transfer relation by the randomness of the input. This is an important fact to keep in mind, for the speed at which a result may be obtained might be greatly increased by using a digital processor, which is slower than a hardware correlator but is capable of normalizing by use of the input function.

Basically, this normalization is the whole point of our study. What we have attempted to make clear is that by choosing the proper *measured normalizing functions* a great deal can be added to the accuracy, speed, and reliability of measurements based on cross correlation and the cross power spectrum. Lest this point appear trivial, the author suggests that there are innumerable situations in which cross correlation is performed on large digital machines when the capability to employ coherence and transfer functions exists at no extra (or less) cost.

REFERENCES

1. Bergland, G. D., "A guided tour of the fast Fourier transform," *IEEE Spectrum*, vol. 6, pp. 41–52, July 1969.
2. Jenkins, G. M., and Watts, D. G., *Spectral Analysis and Its Applications*. San Francisco: Holden-Day, 1968.
3. Bracewell, R., *The Fourier Transform and Its Application*. New York: McGraw-Hill, 1965.
4. Hope, G. S., "Machine identification using fast Fourier transform," *IEEE Paper 71 CP 110-PWR*.
5. Allen, J. B., "Estimation of transfer function using the Fourier transform ratio method," *AIAA J.*, Mar. 1970.
6. Roth, P. R., "How to use the spectrum and coherence function," *Sound and Vibration*, vol. 5, Jan. 1971.
7. Enockson, L. D., "Frequency response functions and coherence functions for multiple input linear systems," NASA Rept. CR-32, Measurement Analysis Corp., Los Angeles, Calif., 1964.
8. Enockson, L. D., "Gaussian approximation to the distribution of sample coherence," U.S. Govt. CFSTI document AD 620 987.
9. Bendat, J. S., and Piersol, A. G., *Measurement and Analysis of Random Data*. New York: Wiley, 1966.

Peter R. Roth (M) is a design engineer with the Santa Clara division of Hewlett-Packard. At Hewlett-Packard, Mr. Roth has been the project leader responsible for the 5450A and 5452A Fourier analyzers and at present is working on the development of measurement procedures that employ the use of digital Fourier analyzers. Before joining Hewlett-Packard in January 1965, Mr. Roth served as an engineering officer with the U.S. Coast Guard for three years. Assigned to the Coast Guard's electronics laboratory in Alexandria, Va., he participated in the development of high-frequency communications and loran A and C equipment design. Mr. Roth received the B.S. and M.S. degrees in electrical engineering from Stanford University and is a member of Tau Beta Pi.