IEEE spectrum

features

A full-function digital multimeter + A lab-quality digital AC voltmeter

...both for $595

HP's new 3469A gives you a general-purpose digital multimeter *plus* a lab-quality digital AC voltmeter—for the price of the AC voltmeter alone. Now, you don't have to buy two (or more) instruments to get the capabilities you need—or compromise on quality to stay within your budget.

As a general-purpose multimeter, the 3469A gives you exceptional capabilities. Its 1 Ω range lets you measure low-resistance components and even contact resistances of a few milliohms, with an accuracy of ±0.25% reading ±0.5% range. To make the low range easily useable, a unique offset adjustment lets you compensate for lead resistance. In the higher ranges (100Ω to 10 MΩ), accuracy is ±0.3% reading ±0.2% range. The 3469A also gives you five DC voltage ranges (100 mV to 1000 V) and six DC ampere ranges (1 μA to 100 mA), with accuracy of ±0.2% reading ±0.2% range or better, depending on range.

As an AC voltmeter, the 3469A is unmatched at any price. You get seven voltage scales, ranging from 1000 V full-scale down to 1 mV full-scale—100 times the sensitivity of other digital meters. You also get a 10 MHz bandwidth capability—100 times greater than other digital multimeters—with a basic accuracy of ±0.3% reading ±0.3% range. And you get a bright, ultra-reliable, shaped-character GaAsP display, that's easier to read than tubes or bar-segment numerals.

Compare the 3469A's specs with any other meter's — and you'll agree that there's no better value, at any price. For further information on the 3469A, contact your local HP field engineer, or write Hewlett-Packard, Palo Alto, California 94304. In Europe: 1217 Meyrin-Geneva, Switzerland.

091/16

HEWLETT hp PACKARD

DIGITAL VOLTMETERS

Circle No. 4 on Reader Service Card

the cover

Pattern for the word "no" as analyzed for the automatic word recognizer described in the article beginning on page 57. The salient features seen in an equivalent sonogram representation of the same word are also present here, thus enabling its ready recognition by the practiced eye

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*

IEEE Publications Operations

IEEE PUBLICATIONS BOARD
C. L. Coates, Jr., *Chairman*
F. S. Barnes, *Vice Chairman*
J. J. Baruch

F. E. Borgnis	R. W. Lucky
R. F. Cotellessa	O. K. Mawardi
David DeWitt	A. A. Oliner
E. K. Gannett	Seymour Okwit
E. E. Grazda	J. E. Rowe
P. E. Green, Jr.	R. L. Schoenfeld
H. E. Koenig	R. H. Tanner

STAFF
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, *Research Manager*
Carl Maier, *Advertising Production Manager*
Marilyn Lee, *Assistant Advertising Production Manager*

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 use the change of address form below. **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*.

MOVING? PLEASE NOTIFY US 4 WEEKS IN ADVANCE

Name

Address

City, State/Country, Zip

(PLEASE PRINT)

Mail to: IEEE Coding Department

345 East 47th Street, New York, N.Y. 10017

ATTACH LABEL HERE

- This notice of address change will apply to all publications to which you subscribe.
- List new address above.
- If you have a question about your subscription place label here and clip this form to your letter.

VOLUME 8 □ NUMBER 8 IEEE spectrum AUGUST 1971

Spectral lines

Why encourage the study of engineering when there is an oversupply of engineers? Some of us believe that because engineers are not fully employed today and the prospects for an early resumption of growth are poor, we should no longer encourage young people to choose our profession. It is easy to understand an emotional basis for that belief but if we look at our situation with the same analytic objectivity that we apply to other people's systems it is possible to conclude that we must continue to display the nature of engineering to the young and to accept those with the aptitudes and the calling.

Most of the nations that now have a substantial engineering profession are committed to a dependence on technology and its improvement. Their institutions do not have the capability for cultural revolutions that can, in a few decades, dedicate and equip their populations for a nontechnological existence and make the necessary population-density adjustments. There are new engineering tasks in satisfying ecological constraints, conserving materials, and providing cultural support equipment for life satisfactions not based on the anticipation of new personal wealth. So some decades from now, when most of us will have finished our years of achievement, an effective engineering profession will still be required.

There are two first-order characteristics of the population of engineers that make the current employment situation an unimportant factor in determining how vigorously we should try to get young people exposed to engineering as a career.

1. Engineering requires both talent and dedication. The level of ability necessary and the strong desire to do engineering are found in a limited fraction of the population of young people. Since engineers constitute only a small percentage of the total working population, the profession should get the best of the willing and able. The way to do this is to present the joys and travails of our work to all young people. There is no reason to conceal the current employment situation but we should not spread the false impression that there will be no future work for engineers. If we do the job honestly and well we will get people with the calling, those who will work hard and creatively to continue the great work begun by our predecessors because it is the thing they want to do and because they have been driven to convert talent into skill by that strong desire. We will not get those looking primarily for secure high wages as a reward for their talent for getting good grades in difficult studies.

2. The rate at which young people should decide to study engineering does not depend on the current employment situation. Roughly, the decision to prepare for the profession is made between the ages of 10 and 20 and an engineer is a full member of the profession between the ages of 30 and 60. The young people to whom we should be talking now are not our competitors for scarce jobs today, they are the engineers our world should have in the '80s and '90s when most of us will be approaching Life Membership in the IEEE.

The conclusion that we should continue to get young people into engineering is independent of a determination of what engineering population will be needed in the future and how many should enter each year. Provision of the right number of people is a serious problem, not for engineering alone, but for all professions and specialties that require extensive training. Equally important and difficult for innovating professions is to determine what to teach that will best prepare people for 30 years of the new and unknown.

In the United States sudden and irremediable unemployment is a damaging experience to every family afflicted. Hence even a small rate of unemployment, say 10 percent, is a major disaster. With the many unknowables in predicting the number of engineers needed decades away, a tempting strategy that can insure full engineering employment at all times is to so restrict entry into the profession that there will surely be a perpetual shortage in the decades to come. That strategy is an abdication of our professional responsibility to society but, more important, it will be corrected by society five to ten years after it is first detected.

As we look decades ahead it seems very plausible that we can propose and institute schemes that provide the world with an adequate supply of people trained in every profession and specialty while protecting those specialists and their families from the indignity and deprivation visited on them today when there is an oversupply of their skills. The pages of this magazine are open for proposals and their discussion. While we are finding ways to make all professions secure from gratuitous indignity let us continue to get the right young people into our profession.

David DeWitt, Editor

Signal processing in acoustic surface-wave devices

Interest in acoustic surface waves is greatly increasing, as evidenced by the growing number of potential applications being investigated. For example, waveguides and amplifiers using the surface-wave principle are considered by many to be precursors of a new broadband ultraminiature circuitry

Gordon S. Kino Stanford University
Herbert Matthews Sperry Rand Research Center

The Rayleigh surface-wave mode propagates on the surface of suitable elastic solids with velocity independent of frequency at about 10^{-5} times the speed of light and an attenuation of about 1.5 dB for 10^4 wavelengths. Thus as a delay line it is far more compact, has much less attenuation, and can be much cheaper than electromagnetic delay lines. Piezoelectric elastic solids permit the launching and detection of the waves by electrodes on the surface. Fabrication technology is similar to that used for monolithic semiconductors and amplification is possible by interaction with carriers in a semiconductor surface layer. Parametric interactions can be used to perform convolution of two signals. The feasibility of deflecting and modulating light beams also has been demonstrated. This rapidly developing art can today produce delay lines, filters, pulse compressors and expanders, and pulse-sequence generators and decoders, and in the future may perform convolution, correlation, and light-beam manipulation in small, economical, low-power structures.

During the final decade of the 19th century, seismologists observed that a series of tremors is transmitted from seismic shock centers. Elastic volume waves transmit the first, and weakest, tremors along a chord through the volume of the earth. The later, dominant tremors are transmitted along the surface of the earth by *elastic surface waves*. The seismic observations were preceded by Rayleigh's theoretical discovery of elastic surface waves in 1885.[1] In his classic paper, Lord Rayleigh predicted the properties of these waves and their seismological importance. He showed that the particle motion associated with an elastic surface wave is largest near the surface of the solid on which the wave propagates, and the amplitude of the motion has components that fall off approximately exponentially into the interior of the medium. He also showed that the wave is nondispersive (i.e., its velocity does not vary with frequency), its phase velocity is somewhat slower than the slowest elastic volume wave in the medium, and its penetration depth into the interior is comparable to a wavelength.

There has been great interest during the past few years in Rayleigh surface waves for use in acoustic delay lines and filters. The interest arises because the wave, being bound to the surface, is accessible and can be readily manipulated. This feature was apparently first recognized in 1963 in a patent for a tapped delay line based on the travel of Rayleigh waves along the surface of a piezoelectric solid.[2] In a piezoelectric material there is an electric field associated with the Rayleigh wave; metal electrodes deposited on the surface can therefore detect the presence of the wave. Reciprocally, if an RF potential is applied between metal electrodes deposited directly on a piezoelectric substrate, Rayleigh waves will be generated. The ease and repeatability of photolithographic fabrication techniques make construction of such acousto-electric couplers relatively simple.

The attenuation of high-frequency acoustic waves in many solids is relatively small. The velocity of an acoustic surface wave on lithium niobate, a piezoelectric material much used for Rayleigh-wave studies and devices at the present time, is about 3.5 km/s. At 200 MHz the wavelength is 17.5 μm and the time delay of a wave is approximately 300 μs/m. The attenuation of such a wave is only about 0.1 dB/cm—that is, approximately 1.5×10^{-4} dB per wavelength. The loss per wavelength at ultrahigh frequencies is thus considerably lower than the loss of an electromagnetic wave in an S-band waveguide. The velocity of the wave is approximately 5 orders of magnitude less than that of light. Consequently, the size of a device constructed for the same frequency range tends to be 5 orders of magnitude less than one based on the use of electromagnetic waves.

The first piezoelectric Rayleigh-wave devices with interdigital transducers, of the type illustrated in Fig. 1, were demonstrated in 1965 by White and Voltmer,[3] who used photolithographic methods to fabricate the transducer pattern. Following this work, a wide variety of surface-wave signal-processing devices and filters have been developed, such as simple- and multiple-tapped delay lines, bandpass and reject filters, and frequency and phase coders and detectors. The filter devices depend for their operation on surface-wave interaction with a specially shaped electrode pattern. Photolithographic techniques are especially suitable for pattern fabrication.

In addition, a large number of devices analogous to conventional electromagnetic-wave devices, but far

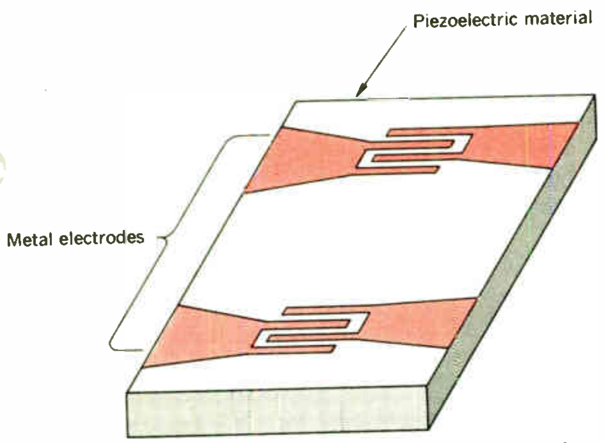


FIGURE 1. Schematic representation of a typical piezoelectric surface-wave device with interdigital electrodes.

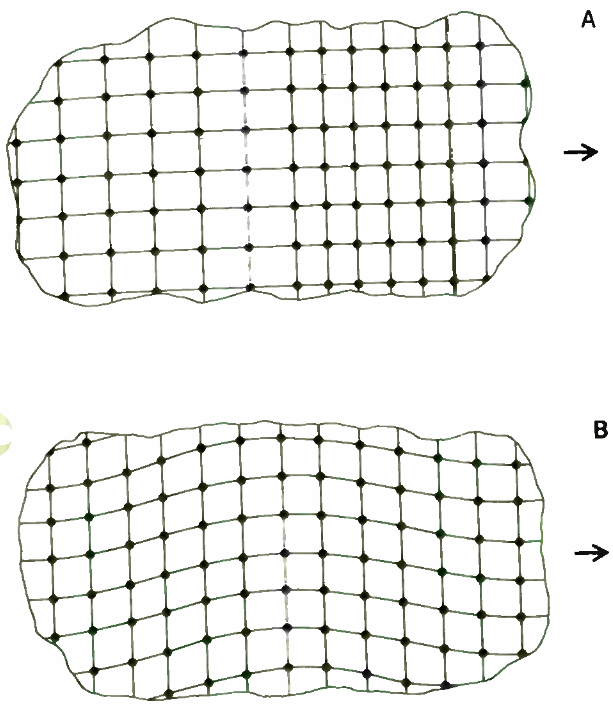
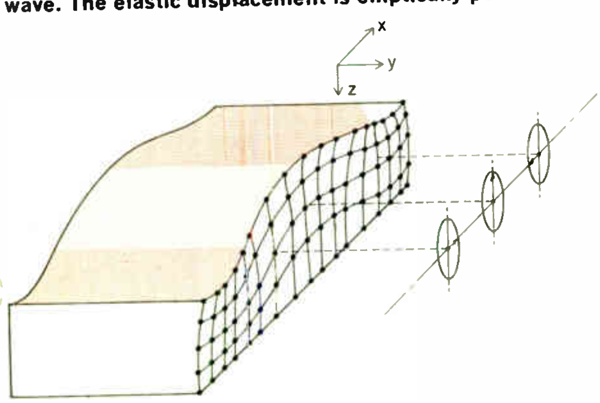


FIGURE 2. A—Pattern showing how a uniform square mesh in a material is displaced from its equilibrium configuration in the presence of a longitudinal wave. B—Similar pattern for a shear wave (transverse wave).

FIGURE 3. Pattern showing how a uniform square mesh in a material is displaced from equilibrium in a Rayleigh wave. The elastic displacement is elliptically polarized.



smaller in size and easily and cheaply fabricated, have been demonstrated. For example, strip waveguides consisting of narrow stripes of dense material deposited on a less-dense substrate have been constructed. In this case, the surface wave is confined to the width of the guide, as well as being concentrated near the surface of the substrate. Strip waveguides have been fashioned into laboratory model power dividers, directional couplers, frequency selectors, and logic devices, as well as dispersive delay devices.[4] Another very interesting and important development is the amplification of surface waves by a mechanism reminiscent of traveling-wave-tube amplification.[5] The waveguides and amplifiers may be precursors of a new broadband ultraminiature circuitry.

Recently a new type of device has been demonstrated that is based on the use of parametric interactions between elastic waves, which are not necessarily but can conveniently be *surface* elastic waves.[6,7] With this process, it is possible to obtain real-time convolution or correlation of two modulated signals, as well as electronically variable time delay and time inversion. This could lead to a wide variety of electronically programmable signal-processing devices.

The exploratory work of the previous five years has produced many interesting new surface-wave studies and developments. There are far too many to describe them all in this article. For further background in the field, the interested reader is directed to an excellent and very complete review paper by R. M. White[8] and to the special issue on microwave acoustics of the IEEE Transactions on Microwave Theory and Techniques.[9]

Waves in a solid

Let us now look at the properties of some of the more important types of elastic waves that can exist in a solid so that we may understand the nature of an acoustic surface wave. The simplest type of elastic wave in a solid is a longitudinal wave in which the material is alternately compressed and expanded. The particle displacements in this type of motion are illustrated in Fig. 2A. A second type of acoustic wave, the transverse wave, is associated with the shear of a material as the acoustic wave passes through it. The particle displacements in this type of motion are shown in Fig. 2B. This wave is slower than a longitudinal wave in most materials.

When the material is of finite cross section and has a free surface, the forces or stress associated with the motion must be zero in a direction normal to the free surface. Because stress is a tensor quantity rather than a vector quantity, certain boundary conditions are more complicated than the equivalent ones for an electromagnetic wave. For example, when an electromagnetic plane wave is incident at an angle to the free surface of an isotropic dielectric, there is a reflected wave with the same polarization. However, if a pure shear or a pure longitudinal elastic wave is incident at an angle to the free surface of a solid, the reflected wave is neither pure shear nor pure longitudinal. Consequently, the boundary conditions at a free surface can be satisfied, in general, only by a combination of shear and longitudinal waves. This combination of wave motions can give rise to a different kind of wave, the Rayleigh wave.

The pattern of elastic displacement in a Rayleigh wave is illustrated in Fig. 3. The wave is traveling in the x direction and the wavefronts are parallel to the y direction.

The figure shows a cross section perpendicular to the surface. The dots represent particles of material displaced from their equilibrium positions, which, in the absence of the wave, are located at equal horizontal and vertical distances. The figure suggests that a Rayleigh wave is the superposition of a shear wave and a longitudinal wave shifted in phase by $\pi/2$. Note, for instance, the highest particle in the cross section. Here the transverse displacement is maximum, the longitudinal zero. As the wave passes, an observer looking from the right at the highest particle would see it trace an elliptical motion clockwise. At a depth of 0.2 wavelength the direction of rotation changes because the longitudinal displacement changes sign. This feature is exhibited in Fig. 4, which shows the depth dependence of the transverse R_t and longitudinal R_l displacements. The curves emphasize the essential characteristic of a surface wave: the disturbance fades away with depth.

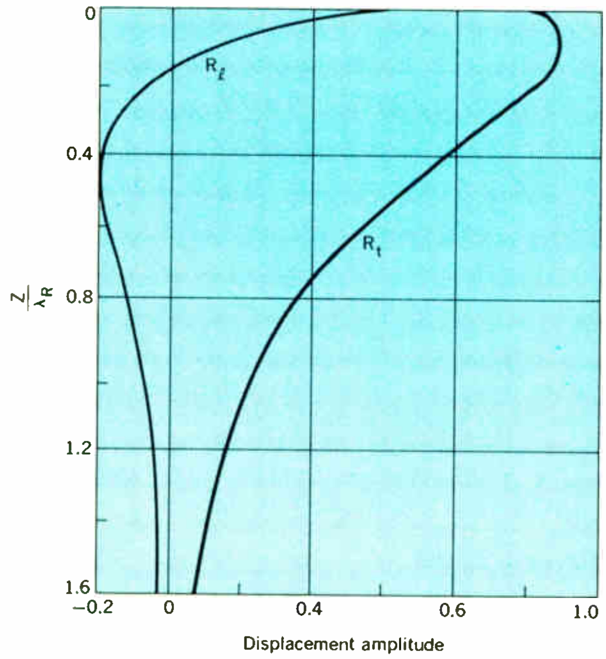
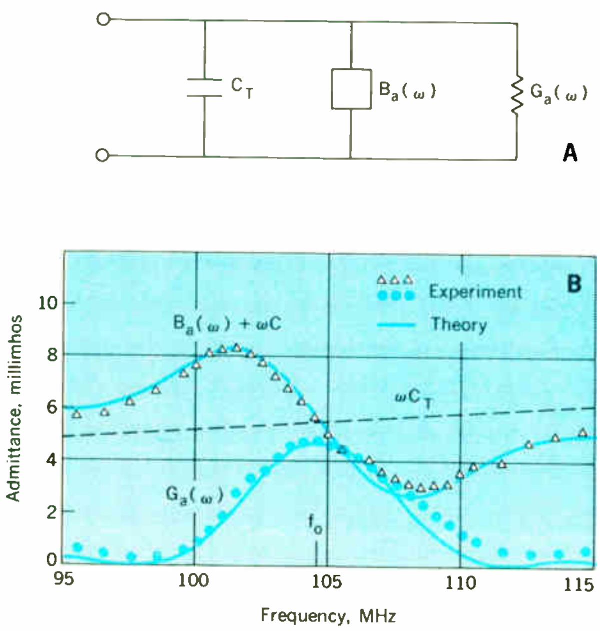


FIGURE 4. Variation of elastic displacement amplitudes as a function of depth for a Rayleigh wave in isotropic material; λ_R denotes the propagating wavelength.

FIGURE 5. A—Shunt equivalent circuit for an interdigital transducer. B—Comparison of theoretical and measured admittance of a five-finger-pair interdigital transducer employing the shunt equivalent circuit.[10]

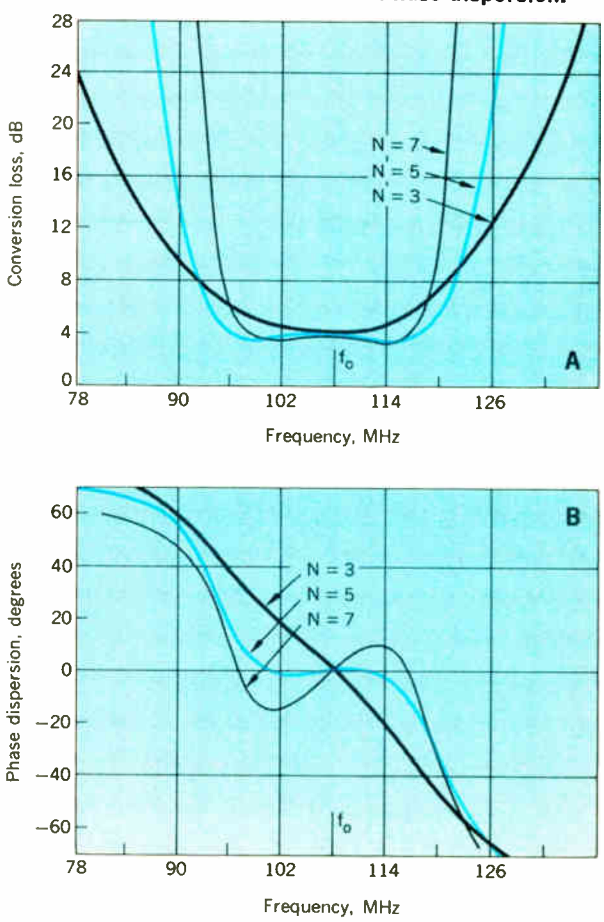


The interdigital transducer

Acoustic surface waves are relatively easy to excite in a piezoelectric material by the use of an interdigital transducer. This is now, by far, the most important type of coupler used. A piezoelectric material is one in which an electric field is generated when a force is applied to it, and conversely. Such materials, to be piezoelectric, must have anisotropic elastic properties; consequently, the behavior of acoustic surface waves propagating in them is more complicated than for an isotropic material. The properties of such waves depend on their direction of propagation relative to the crystalline axes, and have been worked out for the most commonly used crystals. A correct choice of axes is particularly important to obtain optimum piezoelectric coupling—i.e., the largest possible value of electric field at the surface for unit power in the surface wave.

An interdigital transducer, illustrated in Fig. 1, con-

FIGURE 6. Coupling between the electric and one acoustic port for transducers with 3, 5, and 7 interdigital periods. A—Theoretical conversion loss. B—Phase dispersion.[11]



sists of two sets of interleaved metal electrodes, called fingers, deposited on the piezoelectric substrate. To generate a wave an RF potential is applied between the adjacent sets of fingers, which are spaced by a distance equal to one-half wavelength at the transducer design frequency. A typical 100-MHz transducer on $LiNbO_3$ has aluminum fingers 0.2 μm thick by 9 μm wide, with 9-μm gaps. The wave excited by the RF potential between a pair of fingers travels at the surface-wave velocity. By the time the wave arrives midway between the next pair of fingers, the RF excitation potential has reversed sign, and the wave excited by the second pair of fingers will be in phase with the wave from the first pair. Thus the excitation due to the second pair is added to the excitation from the first, and so on. The mechanism is reciprocal, and hence the transducer that excites a wave will also detect it.

These couplers work equally well for waves propagating in either direction, since the couplers are symmetric. Therefore a transducer has a minimum electric-to-elastic conversion loss of 3 dB because half the elastic power is excited in each direction. There is a similar loss in conversion from elastic to electric power so the minimum net terminal loss from electrical input to electrical output is 6 dB. It is a simple matter to make terminations for the signal traveling in the wrong direction by using absorbing material, such as wax or Teflon, placed on the surface of the solid substrate.

An engineering design theory for interdigital transducers, developed in detail by a group at Stanford University, accurately predicts transducer properties.[10,11] The theory is based on an electromechanical equivalent circuit for one pair of fingers. One result of the theory is the shunt representation shown in Fig. 5A. The circuit elements $G_a(\omega)$ and $B_a(\omega)$ represent, respectively, radiation conductance and susceptance. The capacitance between the two finger sets is C_T. The measured admittance of a five-finger-pair transducer is compared with theoretical values in Fig. 5B.

The transducer bandwidth is limited because its properties are much like those of its electromagnetic counterpart, the end-fired antenna array. If there are a large number of fingers, and thus the interdigital line is many wavelengths long, even a slight change from synchronism frequency causes a large mismatch in phase, and net excitation is relatively small. Therefore, a long transducer has a relatively narrow bandwidth.

It has been shown that transducer input conductance $G_a(\omega)$ is given by

$$G_a = G_0 \left(\frac{\sin x}{x}\right)^2 \qquad (1)$$

where G_0 is the radiation conductance at the center frequency ω_0, $x = \pi N[(\omega - \omega_0)/\omega_0]$, and N is the number of finger pairs. The power radiated for a given voltage across the transducer is proportional to the square of the number of fingers, N^2, and the midband radiation conductance G_0 is thus proportional to N^2. The conductance G_0 is also related to the electrical capacitance of the transducer by

$$G_0 = \frac{4}{\pi} k^2 \omega_0 C_T N \qquad (2)$$

where C_T is the transducer capacitance ($C_T \propto N$) and k is the effective piezoelectric coupling coefficient, a quantity proportional to the electric field at the substrate surface when there is unit power flow in the acoustic wave. For the strongest coupling material available, $LiNbO_3$, $k^2 \approx 0.05$ under optimum conditions.

Since the transducer must be driven from a source through a matching network, the frequency response is determined by matching-circuit properties as well as by the properties of the transducer periodic structure. Since the radiation susceptance is comparatively small, the transducer can be well matched to a resistive load by tuning out interfinger capacitance C_T with a simple inductance. In that case, the matching circuit $Q = \omega C_T/G_0 = \pi/4k^2N$. On the other hand, it follows from Eq. (1) that the transducer periodic structure "Q" is about N. Since the transducer bandwidth is limited by the greatest of these Q's, the largest bandwidth is obtained when they are equal—i.e., when $N^2 = \pi/4k^2$. For $LiNbO_3$ the value of N for optimum bandwidth is thus $\sqrt{5\pi}$ ($\approx$ 4). Since both G_0 and C_T are proportional to the finger length, a designer can choose this length to match the midband radiation conductance to the source, and, within limits, the number of fingers that are required to obtain a given bandwidth.

A surface-wave transducer is a three-port device—one electric, two elastic. Figure 6 shows the conversion loss as a function of frequency from a 50-ohm source to one of the two acoustic outputs for three different transducers on a lithium niobate surface. These calculated curves show that for this particular case the use of five finger pairs provides the widest bandwidth and smallest conversion loss.

Many materials have been used in exploratory studies of surface-wave phenomenons and devices. These include cadmium sulfide, zinc oxide, bismuth germanium oxide, lithium tantalate, potassium sodium niobate, crystal quartz, PZT (lead zirconate titanate) ceramic, and lithium niobate, of which the last three are the most widely used. Lithium niobate and PZT are strongly piezoelectric. PZT ceramic is cheap, but dielectric loss and graininess restrict its use to frequencies below 50 MHz. Piezoelectric coupling is comparatively weak in quartz, but this can be an advantage in applications involving complex transducer patterns with large numbers of fingers. Rayleigh waves with a small temperature coefficient of velocity are possible by using a particular plane of quartz,[12] making it an important material for some devices.

The lowest operating frequency of acoustic surface-wave devices is limited entirely by the allowable size. At present the upper operating frequency is limited by fabrication difficulties to about 1 GHz. A typical 500-MHz transducer on quartz has interleaved metal fingers 1.5 μm wide by 3 mm long, separated by 1.5-μm gaps. Usual photolithography techniques can be employed to make transducers up to 600 MHz routinely. Transducers with an operating frequency up to 3.5 GHz have been made by using a scanning electron microscope to expose the photoresist required in the photolithographic method.[13]

Delay-line devices

The surface-wave devices now being readied for use in radar, sonar, communication, and computer apparatus are tapped delay lines. The interdigital transducers are employed for input and output and as taps on the surface-wave delay line. The required signal-processing function is achieved through proper choice of the location, the frequency response, and the coupling of each tap.

The coupling desired from a particular tap is obtained

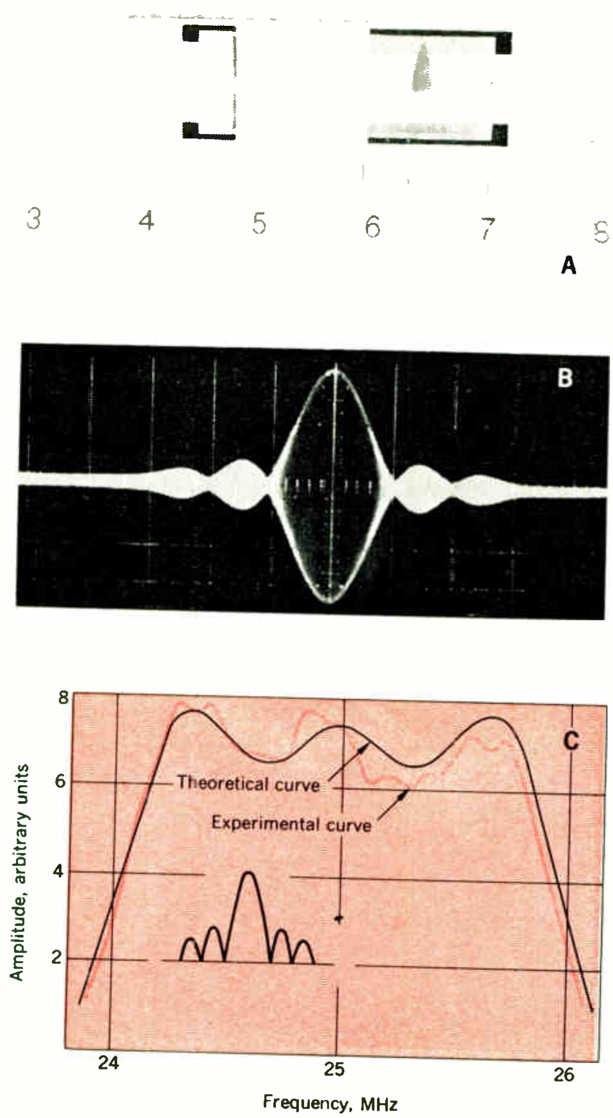


FIGURE 7. A—Bandpass filter with a broadband transducer on the left.[14] Transducer on right has an active filter length that varies as (sin kx)/x with distance. Scale is in centimeters. B—Impulse response of the bandpass filter (amplitude vs. time with a δ-function input). C—Frequency response of the bandpass filter.

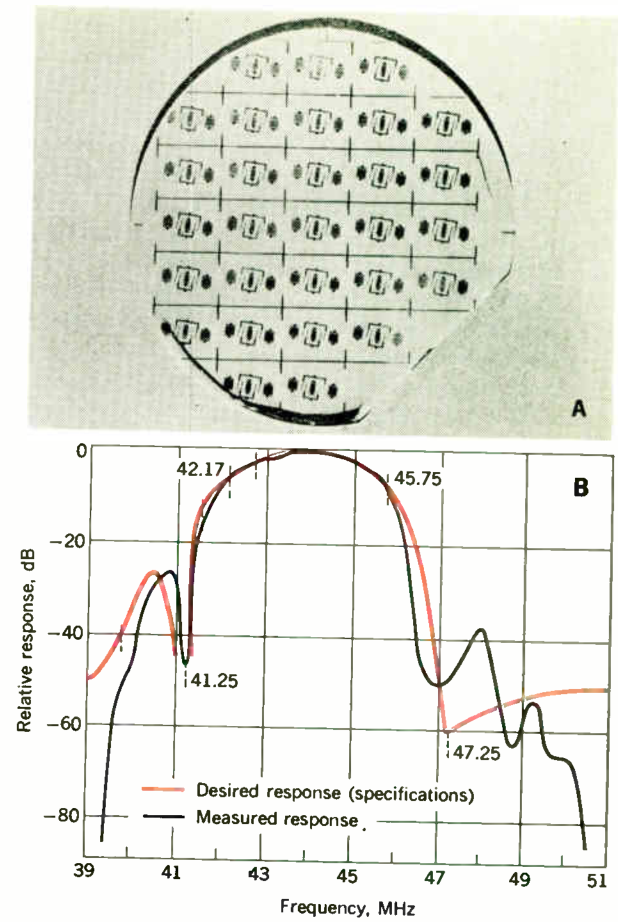
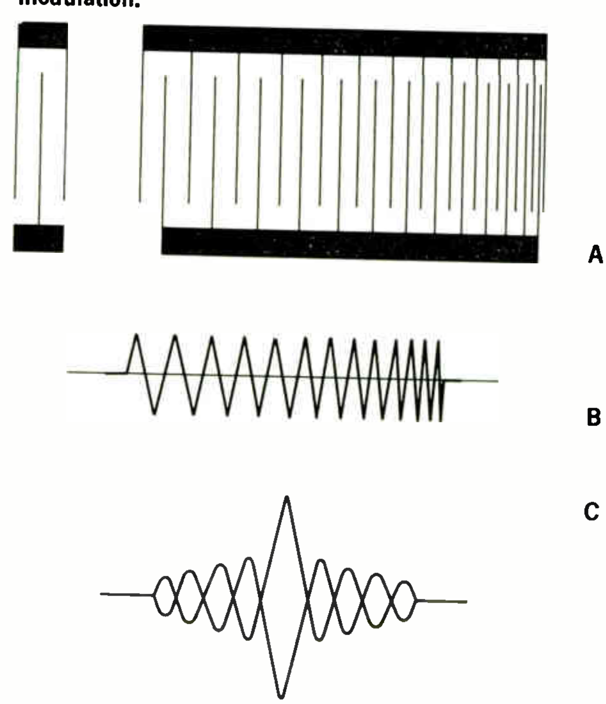


FIGURE 8. A—Photograph of several surface-wave filters for color television IF circuit on a PZT substrate. B—Response curves. (Courtesy A. de Vries, Zenith Microcircuit Facility, Elk Grove, Ill.)

FIGURE 9. Schematic representation (A) of a pulse-compression filter for chirped (i.e., linear frequency-modulated) signals (B). Actual output voltage (C) has the input-pulse center frequency with a (sin αt)/t amplitude modulation.



by choosing the active length of the tap fingers. Figure 7 shows an application of this principle[14] to a 25-MHz aluminum-on-quartz bandpass filter. The input transducer at the left of Fig. 7A has only a few fingers; when it is excited by a voltage pulse of short duration, it radiates a spatially narrow surface-wave pulse. The output transducer on the right has many finger pairs. The active finger length—that is, the amount adjacent fingers overlap—is chosen to vary as (sin kx)/x with distance along the delay line. The oscilloscope trace in Fig. 7B shows the voltage generated as the narrow acoustic pulse moves across the output transducer. This output signal is the impulse response of the filter. As might be expected, it has a (sin αt)/t time dependence. It follows that the frequency response of the filter, which is the Fourier transform of the impulse response, should be almost rectangular. The curves in Fig. 7C compare theoretical and measured frequency response.

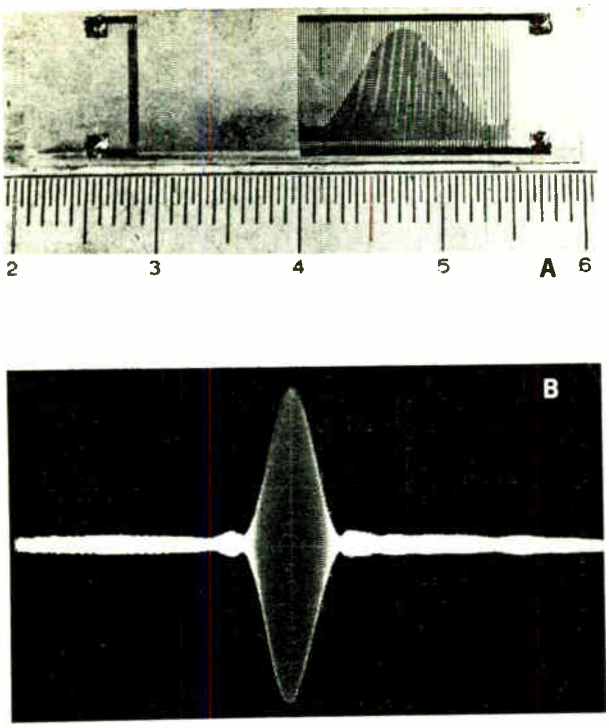
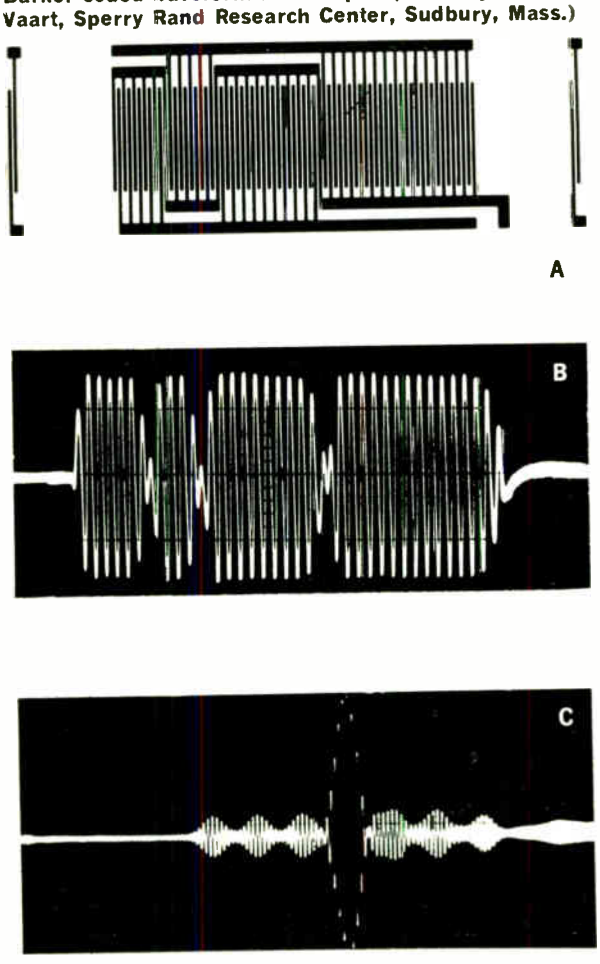


FIGURE 10. Surface-wave FM pulse-compression filter with the tap coupling weighted according to a cosine-squared law to reduce sidelobes in the compressed pulse. Scale is in centimeters.[15]

FIGURE 11. A—Surface-wave tapped delay line for generating and correlating phase-coded waveforms. B—The 7-bit Barker-coded waveform. C—The output with a 7-bit Barker-coded waveform at the input. (Courtesy H. van de Vaart, Sperry Rand Research Center, Sudbury, Mass.)



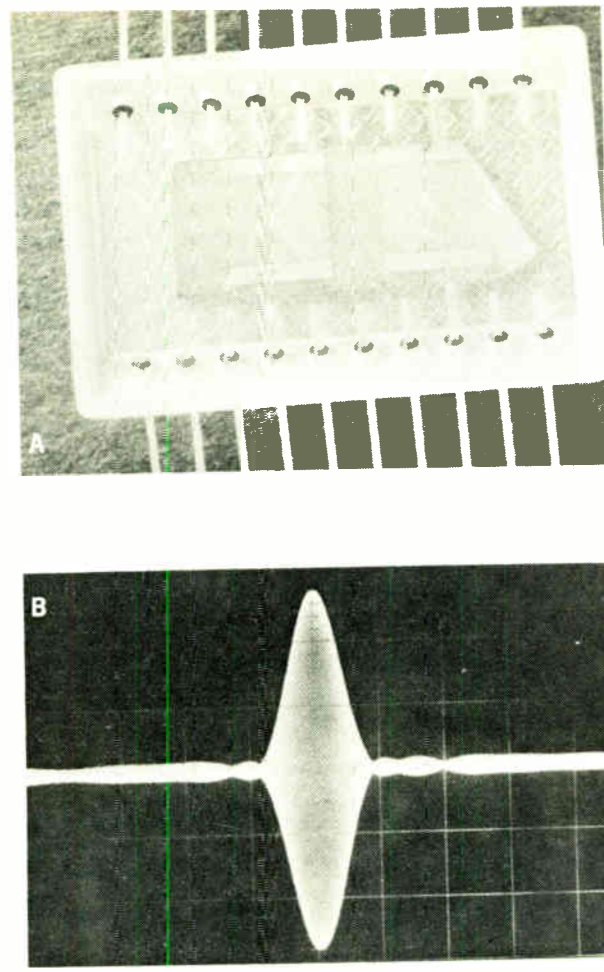


FIGURE 12. Advanced form of FM pulse-compression filter in which weighting of tap coupling is employed. (Courtesy R. Tancrell and M. G. Holland, Raytheon Research Division, Waltham, Mass.)

Similar filters are currently being developed for use as integrated-circuit IF filters in color television receivers. Several transducers are required: one as a passband filter, to sort the video from the sound signal; another as a trap for the sound signal. Filters made for this purpose are shown as Fig. 8A. The input transducer is placed between two output transducers. Extra ground electrodes are placed between the transducers to minimize direct electrical pickup. The transducers are deposited on a PZT substrate. The experimental and theoretical filter response characteristics are shown in Fig. 8B.

Coded waveforms are being increasingly used in both radar and communication systems; surface-wave delay-line devices are very well suited to this application and are being actively developed. Frequency-coded waveforms have been used for many years in sophisticated radars to increase range resolution and discrimination against certain kinds of interference. In the most common form a "chirped" (i.e., linear FM) pulsed carrier signal is radiated and returning echoes are processed in a pulse-compression filter. Much work has been done to develop surface-wave devices for this application.

The device shown schematically in Fig. 9A has a broadband input transducer on the left. The finger-pair spacing in the output transducer is monotonically decreased, so the pairs on the far left are most widely spaced. Thus, the left part of the output transducer responds to lower frequencies and the right to higher frequencies. Suppose a pulse of constant-amplitude, "chirped" RF voltage, in which the frequency decreases linearly with time, as shown in Fig. 9B, is applied to the input transducer. As the surface-wave pulse enters the output transducer from the left-hand side, the low-frequency part of the output transducer does not respond well to the leading high-frequency part of the surface-wave pulse. When the pulse exactly fills the transducer, each finger pair is located at the correct position and has the correct spacing to respond well to the frequency component of the pulse just under it. As the pulse leaves to the right, the high-frequency half of the output transducer does not respond well to the trailing low-frequency part of the surface-wave pulse. The actual output voltage has the input-pulse center frequency with a $(\sin \alpha t)/t$ amplitude modulation, as shown in Fig. 9C. The filter compresses the chirped input pulse to the narrower central lobe of the $(\sin \alpha t)/t$ output.

An improved pulse-compression filter is shown in Fig. 10.[15] The finger overlap, and thus the coupling at each frequency tap, varies according to a cosine-squared law. Such amplitude weighting is a well-known signal-processing technique whose purpose is to reduce sidelobe amplitudes at the expense of a tolerable increase in main-lobe width. Pulse compressions of 100:1 or greater can be obtained by these techniques.

Tapped surface-wave delay lines can be designed to generate and detect phase-coded digital signals. Such lines are expected to be used in sophisticated communication, navigation, and identification applications. When properly excited, the device shown in Fig. 11A, which consists of seven sets of ten finger pairs, will generate a 7-bit surface-wave sequence.[16] Each bit contains 5 RF cycles generated by ten finger pairs when excited by a voltage pulse of ½-cycle duration. The pulse is applied to all pairs simultaneously and seven surface-wave pulses, one for each set of ten pairs, are generated and travel in sequence to the right. (Seven others travel to the left.) The radio frequency of these pulses is determined, of course, by the interfinger distance and the surface-wave velocity. The pulse sequence is the 7-bit Barker code $(+++--+-)$. Since the finger overlap is constant, each bit has the same amplitude. The bit sign is determined by the phase with respect to the common contact bar, which connects half the fingers, as shown in the figure; the remaining half are connected to either of two other contact bars. The sequence generator can thus be excited by a balanced-input network to generate surface waves of opposite phase. In the device illustrated in Fig. 11A there are

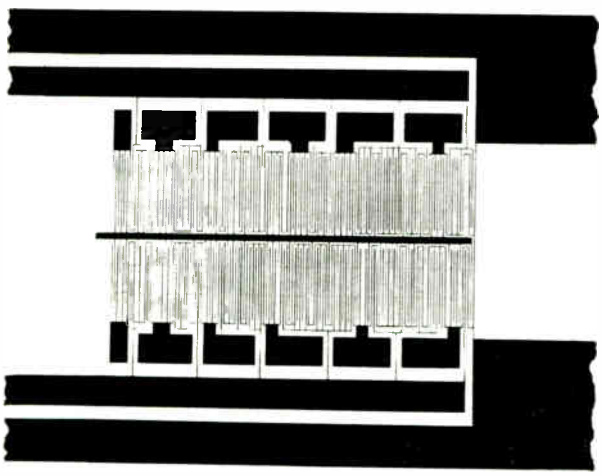


FIGURE 13. Input transducer for delay line using Golay-coded complementary sequences. Upper half of transducer generates Golay-coded waveform; lower half generates complement. Output transducer is identical.

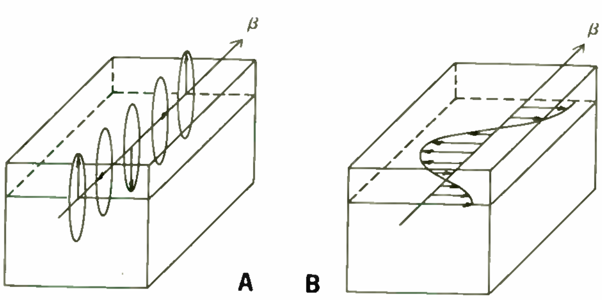


FIGURE 14. Polarization of the elastic displacement for the two major classes of surface waves in layered substrates. A—Generalized Rayleigh waves, elliptically polarized. B—Love waves, transversely polarized.

FIGURE 15. Variation of the elastic displacement amplitude with distance from the surface for the three lowest-frequency Love-wave modes. A—Fundamental. B—First "antisymmetric." C—First "symmetric."

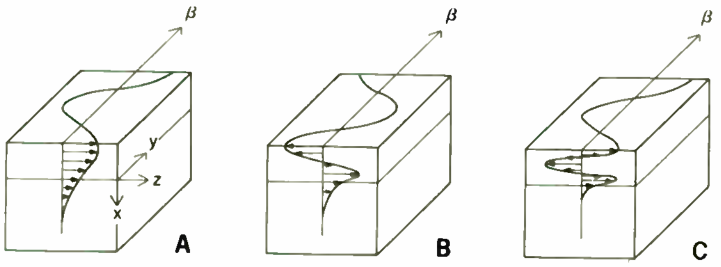
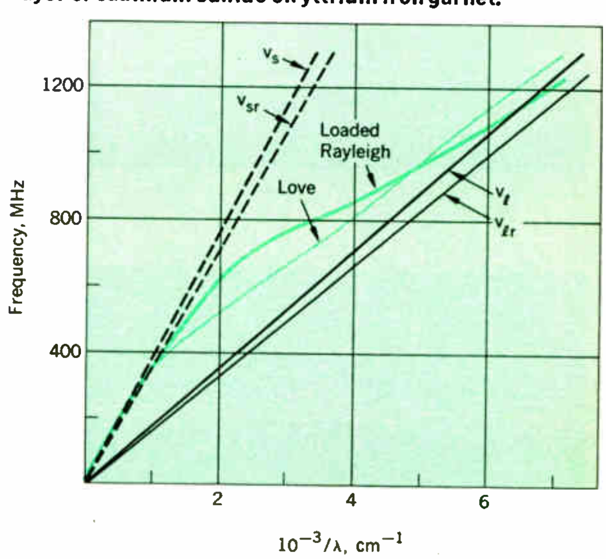


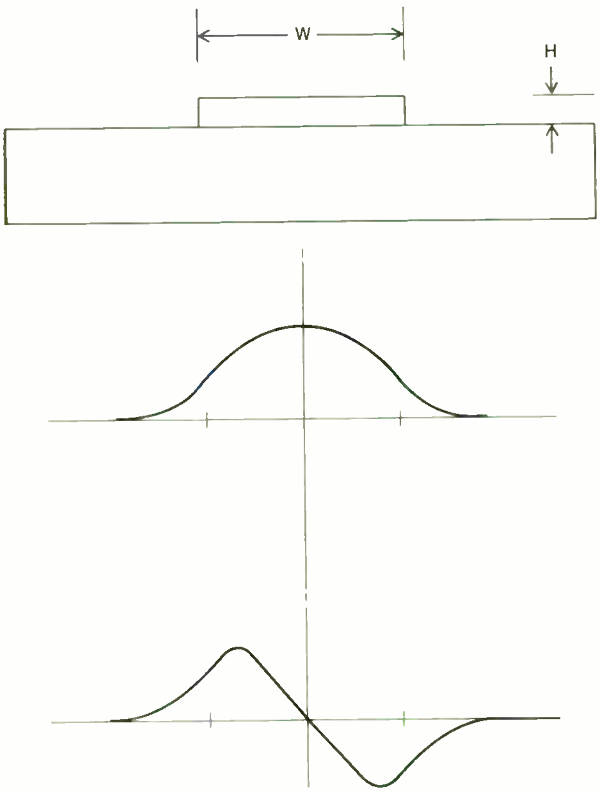
FIGURE 16. Dispersion characteristics of the lowest-order Love-wave and loaded Rayleigh-wave mode in 1-μm layer of cadmium sulfide on yttrium iron garnet.



broadband transducers placed on each side of the sequence generator. The coded output voltage shown in Fig. 11B was taken from the transducer on the right. If this coded voltage is used to excite the left-hand transducer of a similar device, the output (Fig. 11C) from the sequence transducer is the autocorrelation of the original coded voltage. Thus, the device generates and responds only to the coded waveform determined by the coding of the sequence transducer.

The devices just discussed are representative of the tapped-delay-line class of surface-wave devices. They can be used in applications with fairly lenient bandwidth, ripple, or other performance requirements. When wide bandwidth, reduced ripple, lowest insertion loss, or special performance is required it is necessary to use more elegant designs, such as those exhibited in the next two figures. Figure 12 shows a linear FM pulse-compression filter with 240 fingers arranged in two 120-finger mirror-image transducers to increase bandwidth and reduce insertion loss.[17] The finger overlap functions for each half are designed to provide sidelobe suppression by amplitude weighting. The filter is designed to compress a 1-μs, linear, 20-MHz chirp centered on 60 MHz; the compression ratio of 20 is obtained with sidelobes 28 dB below the main-lobe amplitude. An exploratory transducer for recirculating memory or DELTIC applications is shown in Fig. 13.[16] The transducer generates two 64-element Golay complementary sequences when excited by a ½-cycle voltage. The sequences are detected by a similar transducer and combined into a single voltage output. Coding allows surface area to be used efficiently.

FIGURE 17. Surface amplitude of the transverse component of elastic displacement in a loaded Rayleigh wave guided by a strip. The amplitude tends to zero with depth from the surface. Two lowest-order modes are depicted.



Bandwidth and a good match to the source impedance are obtained by a coded transducer stretched out along the propagation path rather than by a few-fingered transducer with considerable length at right angles to the propagation path. Golay coding is preferred because the correlation response of these complementary sequences has theoretically zero sidelobes.

Surface waves on layered substrates

We have shown in Fig. 9 how a dispersive delay line is constructed using special types of transducers. Alternatively, the normal Rayleigh wave can be made dispersive by depositing a layer of different material on the substrate, as shown in Fig. 14A. A surface wave will propagate in the layered medium, its penetration depth into the substrate being small at very high frequencies and large at low frequencies. Consider, for example, a wave traveling in a gold film deposited on $LiNbO_3$. The surface-wave velocity in gold is much lower than in $LiNbO_3$. The wave velocity will be low at high frequencies since most of the acoustic energy is in the gold layer. At low frequencies, since most of the energy is in the $LiNbO_3$, the wave velocity is high. The properties of elastic surface waves in layered substrates are basic to distributed-wave devices, such as strip waveguides. In addition, the next generation of tapped-delay-line devices will probably use layered substrates not only to exploit the additional flexibility provided by combining materials with complementary properties, but also to avoid, when possible, the use of expensive piezoelectric substrate materials.

Two distinct classes of elastic surface waves, distinguished by the polarizations of their elastic displacement vectors, can exist in a structure consisting of a substrate with a thin overlay. The first type, already mentioned and illustrated in Fig. 14A, is a generalization of the familiar Rayleigh wave. The displacement vector, which is elliptically polarized, has two components, one in the direction of wave propagation, the other normal to the surface. In the other class, Love waves (illustrated in Fig. 14B), there is only one displacement component, parallel to the surface and perpendicular to the propagation direction.

A thin layer on a substrate is a mechanically loaded plate and, since a plate is an elastic waveguide and can propagate a great variety of modes, it is not surprising that there are a large number of generalized Rayleigh-wave and Love-wave modes. The elastic displacement amplitudes for the lowest-order and two higher-order Love-wave modes are shown schematically in Fig. 15. The illustrations emphasize the correspondence to the three lowest-frequency horizontal shear modes of a thin plate. A similar correspondence exists between generalized Rayleigh-wave modes and Rayleigh–Lamb modes of a thin plate.[18]

Generalized Rayleigh waves, having two components of elastic displacement, are physically more complex than Love waves. Two discrete types of generalized Rayleigh waves are possible. For one type the layer is stiffer and less dense than the substrate, and the phase velocity of the generalized Rayleigh wave *increases* with increasing propagation constant up to the point where the wave velocity becomes equal to the shear-wave velocity in the substrate.[19] The layer is said to "stiffen" the substrate. An example is aluminum on T-40 glass. For the other type, in which the layer is less stiff and more dense than the substrate, the phase velocity *decreases* with increasing propagation constant. Such a layer is said to "load" the substrate. An example is gold on $LiNbO_3$. Love waves do not exist in stiffened substrates, but an unlayered substrate will support a type of Love wave if the substrate is piezoelectric or magnetostrictive.[20–22]

The lowest-order generalized Rayleigh-wave and Love-wave modes propagate at frequencies approaching zero, but the higher-order modes are cut off below a minimum frequency. The cutoff frequency increases as layer thickness decreases, and at about 1 μm the lowest cutoff frequency is of the order of 1 GHz for most materials.

The propagation properties of loaded Rayleigh and Love waves are very similar. Dispersion curves for the lowest-frequency modes in a typical material pair are compared in Fig. 16. At high frequencies very little elastic energy is carried by the substrate, and the layer properties dominate the wave propagation. At high frequencies the Love-wave velocity approaches the velocity of shear waves in the layer, whereas the loaded Rayleigh-wave velocity approaches the velocity of free-surface Rayleigh waves in the layer. At low frequencies appreciable elastic energy is carried by the substrate, and substrate properties dominate the wave propagation. The frequency at which the phase velocity is slowed appreciably compared with the low-frequency asymptotic velocity depends on the layer thickness. When the layer is very thin compared with a propagating wavelength, the slowing of low-frequency waves is almost negligible.

Loaded Rayleigh-wave or Love-wave dispersion can be used to make a pulse-compression device based on the fact that the group velocity of low-frequency waves is greater than that of high-frequency waves.[23] These layered-substrate pulse compressors can operate over large fractional bandwidths, and are presently used in RF applications. In the UHF range, layers a few micrometers thick are required.

A strip of layer material that loads a substrate will act as a waveguide. The velocity of a loaded Rayleigh wave in the strip is slower than the velocity of the free-surface waves on each side, so a wavefront is bent toward the strip and the strip acts as a waveguide in the same manner as a dielectric guide. If a layer material on a substrate stiffens it, a waveguide mode can propagate in a strip of unlayered surface strips because as the wave velocity is lower along this strip, waves will tend to be confined there. The configuration of the two lowest

FIGURE 18. Two types of Rayleigh-wave amplifier. A—Amplifier using a semiconducting or photoconductive piezoelectric substrate, such as cadmium sulfide.[26] B—A separated-medium amplifier.[5]

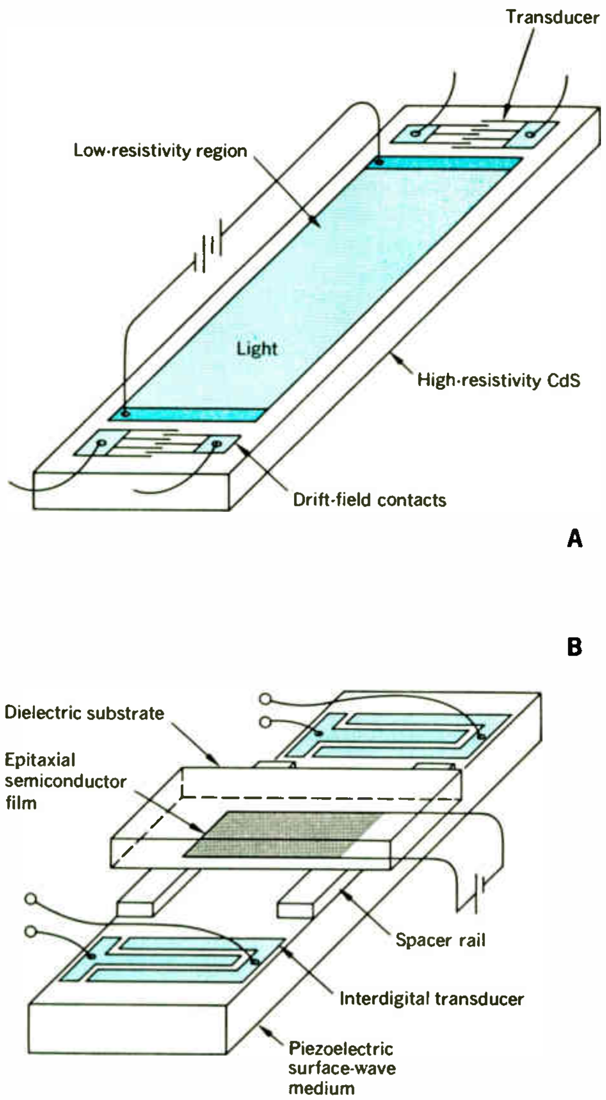
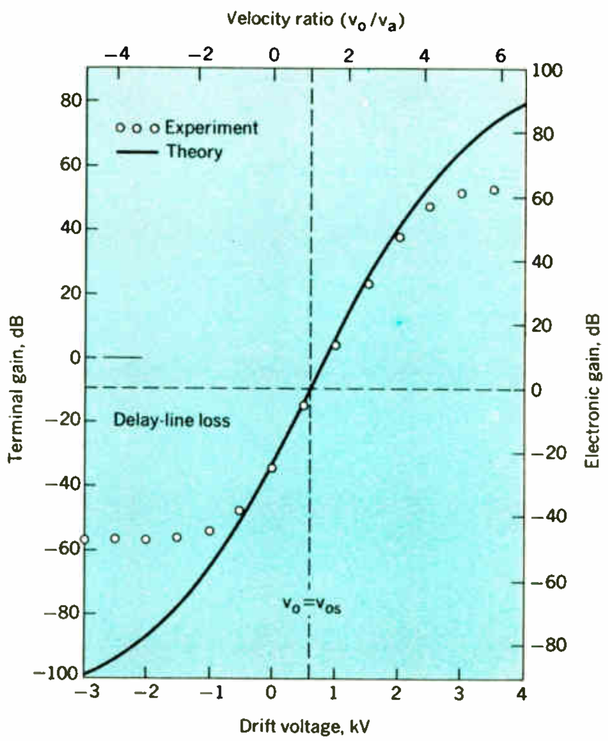


FIGURE 19. Measured and theoretical gain of a typical electric surface-wave amplifier, where v_0 denotes the electron drift velocity and v_a the Rayleigh velocity. The theory was calculated with f = 106 MHz, gap spacing h = 600 Å, semiconductor thickness d = 1 μm, semiconductor mobility μ = 650 $cm^2 \cdot V^{-1} \cdot s^{-1}$, semiconductor length ℓ = 1 cm, and σd = 10.3 μmho, where σ is the semiconductor conductivity. (Courtesy T. Reeder and E. Westbrook, Hansen Laboratory of Physics, Stanford University.)



modes in strip waveguide is sketched in Fig. 17. The amplitude of the elastic displacement at the strip boundary is not zero but extends to each side, and energy can be transferred to an adjacent strip.[24] This is a basis for designing surface-wave "components" for an ultraminiature elastic-wave circuitry. The wave slowing or speeding due to a layered substrate can be used as a basis for designing structures analogous to optical lenses.[25]

At present the only economically feasible means for depositing layers—pyrolytic deposition, sputtering, vacuum evaporation, etc.—yield polycrystalline layers. At ultrahigh frequencies, wave scattering by grain boundaries tends to cause excessive propagation losses. This restraint on the practicality of layered substrate devices will be removed if expected developments in heteroepitaxy technology are realized.

Amplifiers

Acoustic-delay-line filters must inevitably have some loss; when delay lines are designed for very long delays or for very high frequencies, the losses can become prohibitive. In addition, there can be difficulties caused by reflection of waves from the transducers. A wave reflected from the output can be rereflected from the input transducer, and so after three transits through the device it will appear at the output. This phenomenon is known as *triple-transit echo.* Typically it is difficult to obtain a triple-transit echo 30 dB lower than the required signal. For communication systems and radar purposes a better triple-transit-echo figure than this is usually required.

For these reasons, it is very convenient to be able to construct a nonreciprocal acoustic delay line with amplification. The use of an internal amplifier rather than an external transistor amplifier gives the additional advantage that the dynamic range of the system is increased. The minimum RF amplitude of very weak signals is limited by the noise of the delay-line–amplifier system. The maximum RF amplitude is limited by the saturation power of the surface-wave devices—that is, the power at which nonlinearity begins. With an internal amplifier it is possible to use the full dynamic range of the system. With an external amplifier the dynamic range is decreased by the delay-line loss.

Acoustic amplifiers that are effectively the solid-state equivalent of a traveling-wave tube have been developed for use with acoustic volume waves. However, it has been difficult to construct reliable amplifiers with net terminal gains. This has proved to be much easier with acoustic surface-wave amplifiers because of the relatively low loss of the transducers and the relatively small interference from competing modes, which are not well coupled to the semiconductor.

White and Voltmer first demonstrated a Rayleigh-wave amplifier of the type shown in Fig. 18A using cadmium sulfide as the substrate material.[26] Cadmium sulfide is piezoelectric; it is also photoconductive, so by illuminating it with light, carriers can be produced near the surface. These carriers can interact with the electric field of a Rayleigh wave propagating along the surface. If a potential is applied across the illuminated region, the carriers can be caused to drift in the direction of the propagation of the waves. It can be shown that if the carrier velocity is larger than the velocity of the acoustic wave, the carriers will deliver energy to the wave and the wave amplitude will increase with distance. On the other hand, if the carrier velocity is less than the velocity of the acoustic wave or in the opposite direction, the wave will be attenuated.

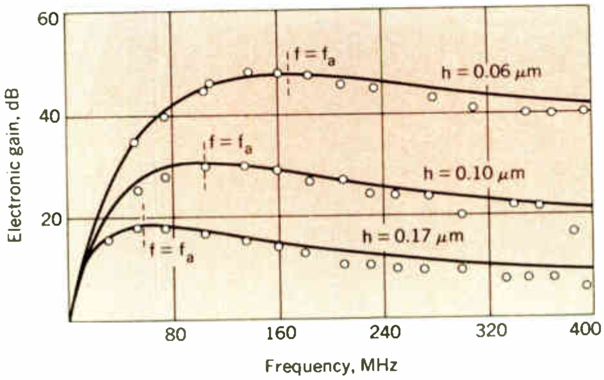
A later development was to separate the functions of the semiconductor and the acoustic material and to use a thin film (1 μm) of silicon deposited on sapphire as the semiconductor.[5] The silicon film was placed close to the acoustic medium, lithium niobate, and separated from it by means of SiO rails, as illustrated in Fig. 18B, the spacings being typically of the order of 500 Å. As with the CdS amplifier, gain is obtained when the voltage applied to drift the carriers makes the carrier velocity greater than the velocity of the waves. Conversely, additional loss is introduced when the velocity of the carriers is less than the velocity of the acoustic wave or is in the opposite direction.

The frequency for maximum gain of this amplifier depends on the gap spacing because the electric fields associated with the acoustic wave decrease as the distance y from the surface of the piezoelectric material as $\exp(-2\pi y/\lambda)$. The frequency for maximum gain is given by the relation $2\pi h/\lambda_R = \epsilon_0/\epsilon_p$, where ϵ_p is the effective dielectric constant of the acoustic material (for example, $\epsilon_p = 50\epsilon_0$ for $LiNbO_3$), λ_R is the wavelength of the acoustic wave, and h is the gap spacing. This implies that for $LiNbO_3$ at 200 MHz the gap spacing for maximum gain is 500 Å.

An experimental curve for gain as a function of drift velocity is shown in Fig. 19. A number of curves for gain as a function of frequency taken with different spacing are given in Fig. 20. Measured gains are as high as 50 dB, and the agreement between theory and experiment is excellent.[27]

These amplifiers exhibit loss in the backward direction, so they can be designed to give good triple-transit suppression. The measured noise figures of such amplifiers have been as low as 5 dB, and the saturation output power, which depends on the spacing between the semiconductor and the acoustic material, as high as 1 watt. The noise figure tends to decrease when the spacing between the semiconductor and the acoustic medium is increased, because the noise is generated in the semiconductor. In addition, higher saturation output power

FIGURE 20. Measured and theoretical gain of a typical surface-wave amplifier, where h denotes gap distance. The theory was calculated wth $v_0/v_a = 3$, $\mu = 606$ cm$^2 \cdot$V$^{-1} \cdot$s^{-1}, $\sigma d = 9.7$ μmho, $\ell = 1.12$ cm. The notation is the same as that employed in Fig. 19.



can be obtained when the semiconductor is placed a large distance from the acoustic material, for the saturation occurs in the semiconductor. Thus the acoustic power can be relatively large when the field at the semiconductor is made relatively small. In both cases, the price paid for high power or low noise is low gain. It is possible, however, to design and construct amplifiers with tapered spacing to get the best of both worlds. It should be possible to design amplifiers with efficiencies of about 10 percent with relatively low noise levels.

At the present time, most of the predictions made about these amplifiers have proved to be accurate and their operation is well understood. The present form of separated-medium amplifier, though a very useful laboratory device, is not necessarily a practical one because of the fine mechanical tolerance required in its construction. However, monolithic amplifiers have been constructed in which a semiconductor such as InSb is deposited directly on the lithium niobate, and have exhibited net terminal gain at 660 MHz.[28] Other forms of monolithic amplifiers are possible in which a piezoelectric material is deposited directly on the semiconductor, or the piezoelectric material is also a semiconductor, as with the CdS amplifier already described. Which of these devices ultimately will prove to be the most useful is not yet clear, because it is not evident which type of material technology will ultimately prove to be the most successful. It is clear, however, that eventually it will be possible to construct a practical amplifier useful in the frequency range from 50 MHz to frequencies greater than 1 GHz.

Parametric interactions

We have so far assumed that the signals used in our surface-wave devices are well below the level of power saturation in the acoustic medium. The medium might be expected to become nonlinear at the point where its response does not obey Hooke's law; i.e., the stress is no longer proportional to the strain. Such effects become noticeable when the strain of the medium (that is, its relative expansion) is of the order of 10^{-5} to 10^{-4}. These nonlinear effects lead to parametric interactions between different RF signals. The interactions have been used with acoustic volume waves to obtain parametric amplification and harmonic generation.

The same type of nonlinearity can be observed with acoustic surface waves. If we consider, for instance, an acoustic surface wave at 100 MHz with a beam width of 1 mm and a penetration depth of 5 μm, a strain of the order of 10^{-3} can be obtained at a power level of approximately 1 watt. Devices of this kind begin to saturate at power levels of a few watts.[29]

Consider the interaction of two oppositely directed waves with RF components that vary as

$$\exp j(\omega_1 t - k_1 x) \quad \text{and} \quad \exp j(\omega_2 t + k_2 x)$$

respectively. If there is any nonlinear interaction present, a strain component will be generated that varies as the product of these two signals—that is, as

$$\exp j[(\omega_1 + \omega_2)t - (k_1 - k_2)x]$$

at a frequency $\omega_3 = \omega_1 + \omega_2$. Associated with this component of strain at frequency ω_3 is an electric field

FIGURE 21. Transducer configurations used for harmonic generation, frequency mixing, and obtaining convolution between two signals. A—Pair of metal electrodes used as the output transducer for second harmonic. B—Relatively coarse interdigital transducer used for the sum frequency output at $\omega = \omega_3$.

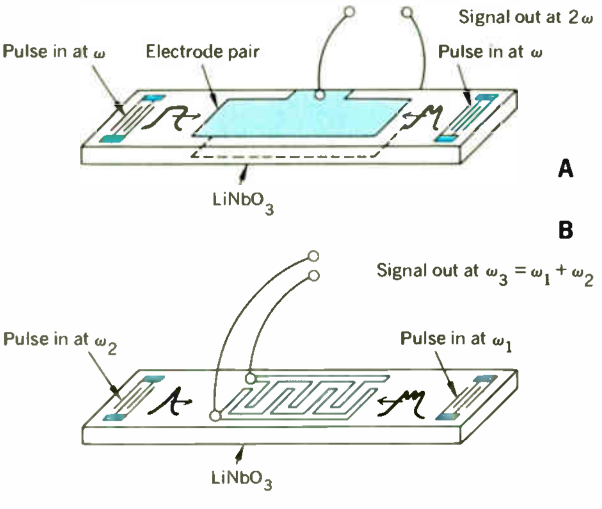
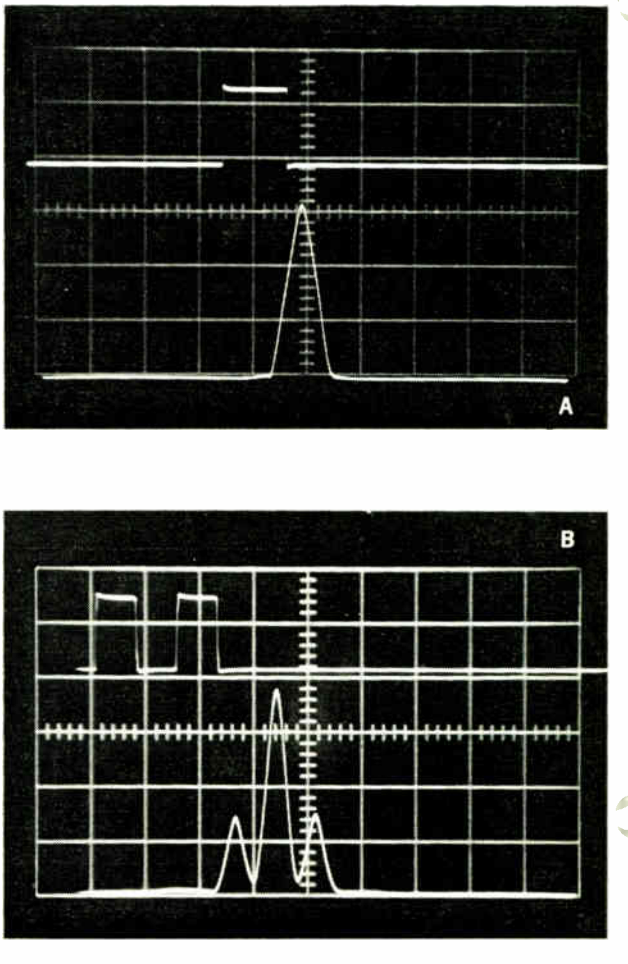


FIGURE 22. A—Autoconvolution of a rectangular pulse observed using a surface-wave convolver with an interdigital output transducer. Scale: 5 μs/div. (Courtesy W. Shreve, Hansen Laboratory of Physics, Stanford University.) B—Autoconvolotion of a double pulse, with the same configuration and frequencies. Scale: 2 μs/div.



with the same variation with time and distance—i.e., a field varying with x as $\exp(-jk_3x)$, where $k_3 = k_1 - k_2$. For the special case in which $\omega_1 = \omega_2 = \omega$, $k_1 = k_2$, and thus $\omega_3 = 2\omega$ and $k_3 \neq 0$; therefore the electric field at the second harmonic is spatially uniform. This kind of operation was first demonstrated in a fundamental experiment by Svaasand[30] using a configuration similar to that shown in Fig. 21A, which employs an output circuit consisting of two metal films on each side of the surface-wave delay line. In the general case with propagation constant $k_3 \neq 0$, the output can be detected on a coarse interdigital transducer with a finger-pair spacing l such that $k_3l = 2\pi$, as shown in Fig. 21B.

Recently, Quate and Thompson showed that if the two input RF signals are modulated, the modulation of the output signal is a convolution of the modulations of the two input signals.[6] They demonstrated this important new application of parametric interactions using elastic volume waves. This and related applications were later demonstrated using acoustic surface waves by Luukkala and Kino.[7] Consider the interaction between two modulated signals

$$F(t) \exp j\omega_1 t \qquad \text{and} \qquad G(t) \exp j\omega_2 t$$

respectively, passing in opposite directions through an acoustic medium. The modulation of the parametric product signal at any point x within the device would be of the form

$$KF\left(t - \frac{x}{v_a}\right) G\left(t + \frac{x}{v_a}\right)$$

where v_a is the velocity of the acoustic wave and K is a coupling constant. The output transducer detects this product signal over a region of length L, and an output signal is obtained of the form

$$H(t) = K \int_{-L/2}^{L/2} F\left(t - \frac{x}{v_a}\right) G\left(t + \frac{x}{v_a}\right) dz \qquad (3)$$

If each signal consists of a pulse whose length is less than the transit time under the detecting transducer, it is then possible to regard the output transducer as being of infinite length. The transformation $t - (x/v_a) = \tau$ shows that the output obtained is

$$H(t) = -Kv_a \int_{-\infty}^{\infty} F(\tau)\, G(2t - \tau)\, d\tau \qquad (4)$$

The output signal may therefore be regarded as a type of convolution of the two input signals. The result of the parametric interaction is analogous to passing a signal through an amplifier. The output obtained is the convolution of the time response of the amplifier and of the input signal. The same analogy can also be applied to passing a signal through a fixed filter, such as a specially shaped interdigital transducer. For parametric interaction the convolution is between an input signal and a reference signal, rather than between an input signal and an amplifier or transducer with a fixed response.

If both the input signal and the reference signal are rectangular pulses the convolution output signal will be in the form of a triangle. Recent results obtained using RF signals of 222 MHz and 188 MHz, respectively, at a level of 0.1 watt, approximately 3 μs long with the output center frequency at 410 MHz at a level of −48 dBm, are shown in Fig. 22A and B.

This technique for obtaining convolution is very useful because one device is capable of carrying out a large number of sorting functions. For instance, coded signals could be used in a manner similar to that already described for coded transducers. However, now a coded reference and a coded input signal are used, and the reference signal can be electronically varied.

Experiments have been carried out in which the reference signal is a δ-function or "pip." The output obtained is the product of a δ-function and the input pulse. Since the reference and input pulses pass each other at twice the acoustic velocity, the output is the original pulse divided in time by a factor of 2. It has also been shown that by applying the δ-function reference signal at different times, the time delay of the output pulse can be varied.

The communications engineer is usually concerned with the correlation between two signals rather than their convolution. He is concerned with the function

$$H(t) = \int_{-\infty}^{\infty} F(\tau)\, G(t + \tau)\, d\tau \qquad (5)$$

By comparing Eq. (5) with (4), it will be seen that a convolution device can only be used to obtain correlation if one of the signals, $F(\tau)$ or $G(\tau)$, is inverted in time.

Consider the configuration shown in Fig. 21B. If a

FIGURE 23. Sketch of the configuration used to obtain time reversal of an input RF pulse.

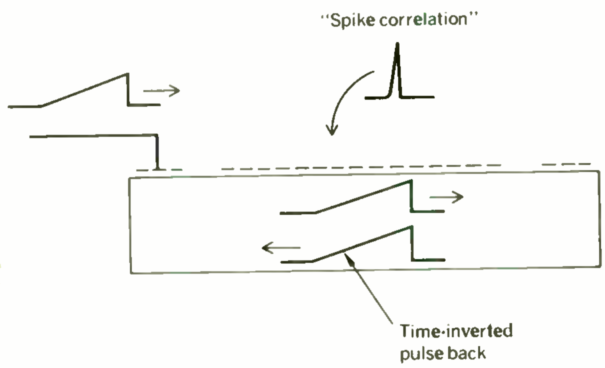
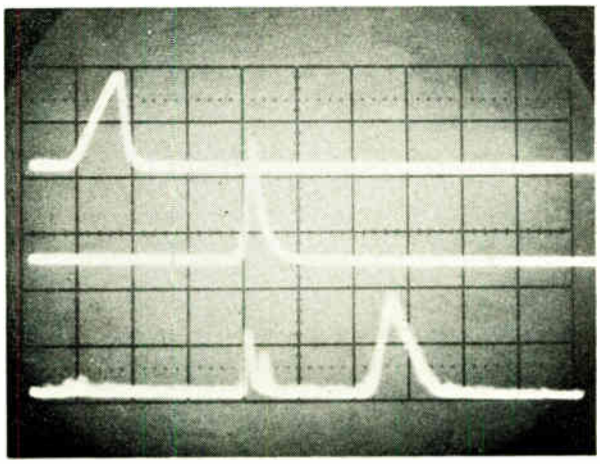


FIGURE 24. Oscilloscope traces of time inversion. From top to bottom: input signal, pump spike (0.2 μs), and time-inverted signal. Scale: 2 μs/div.



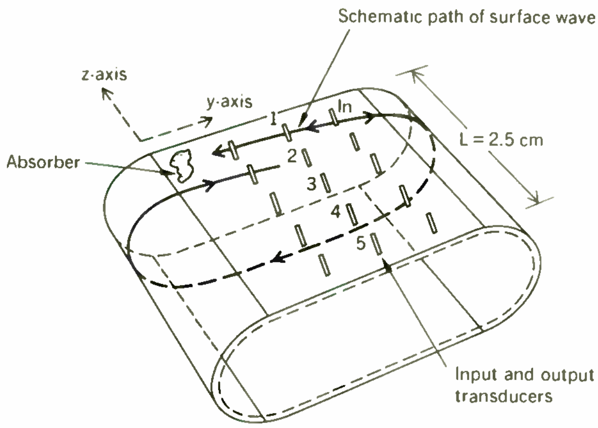


FIGURE 25. Method used to obtain long time delays by passing a wave several times around a piece of $LiNbO_3$ with rounded ends. Transducer labeled "In" is input and those labeled 2–5 are used as outputs.[32]

signal of frequency ω_1 is inserted on the right transducer and a signal of frequency ω_3 is inserted on the center transducer it is possible to obtain a return signal or idler out of the left transducer at frequency ω_2.

Let us assume the input signal at ω_1 to be an asymmetrical pulse, as shown in Fig. 23, and the signal on the center transducer to be a pip. When the asymmetrical pulse arrives under the center transducer, an idler pulse will be generated that is the product of the two signals. Since, however, generation occurs directly under the transducer, the tail end of the original pulse will arrive back at the original input transducer first, and the front part of the pulse will arrive back at the tranducer last; i.e., the output pulse will be a time-inverted version of the input pulse.

An experiment has been carried out by Luukkala and Kino[7] using an input RF signal at 210 MHz, a pip on the middle transducer with a carrier frequency of 450 MHz, and an output at 290 MHz. It will be seen from the oscilloscope trace shown in Fig. 24 that a time-reversed output was obtained. By the use of such a device in combination with a convolution device, correlation between two signals can be obtained.

There are a wide variety of possible devices based on these parametric principles. They have great advantage in being electronically variable. In addition, they provide real-time processing; that is, the time–bandwidth product is already about 75. The results obtained compare favorably with the rate at which real-time determination of the convolution of two such signals could be carried out in a large-scale computer. Here, by the use of acoustic processing, the results are achieved in a relatively small crystal. It is expected that in the future these kinds of devices will exhibit much larger time–bandwidth products and can be used for other kinds of processing.

Similar principles can be used to make parametric amplifiers and oscillators, but useful devices have not yet been demonstrated. A related device that has been demonstrated makes use of the characteristics of an interdigital transducer for the parametric pumping.[31] When the capacitance across an interdigital transducer on $LiNbO_3$ is varied, the velocity of a wave passing underneath it can be varied by as much as 2 percent, which is a relatively large perturbation. A varactor connected across a transducer and pumped at approximately twice the center frequency of the transducer is equivalent to altering the impedance of the transducer at twice its normal frequency and thus provides a parametric pumping action. Such a device, when pumped with power levels of the order of a few milliwatts, broke into oscillation at a frequency near the transducer frequency. At lower power levels, amplification of an acoustic signal passing under the transducer could be obtained.

Conclusions

We have discussed a wide variety of devices that are based on acoustic surface waves. At the present time the field is very active, and there are still an exciting number of new concepts being studied in the laboratory. One exciting possibility is that of storing a complete television frame on one acoustic delay line. This would require storage times of the order of 16 ms. If this could be done, one TV frame could be compared with the previous one, and the difference signal used. Since there is relatively little change from one frame to the other, the amount of information required to be transmitted would be far less, and hence the bandwidth required would be decreased by factors of as much as 10. There is thus a great incentive to make very long delay lines.

One way of doing this is to make a spiral waveguide circuit.[32] Another is to send a signal around and around the line. One configuration being used for this purpose is shown in Fig. 25.[33,34] In this case, the wave would follow a helical path. At the present time, a signal has been sent in a helical path around a device of bismuth germanium oxide. Delays of as much as a millisecond have been obtained this way. By using an internal acoustic amplifier to make up for the losses, milliseconds of delay have been obtained in a similar device in which the acoustic wave repetitively passes the same transducer.

Another need is for low-cost nonpiezoelectric material for the delay line. The problem is obtaining transducers. One approach is to lay, by vacuum deposition or other means, a piezoelectric material on top of nonpiezoelectric material and place an electrode structure either on top of the piezoelectric material or in the interface between the two mediums. Such transducers have been demonstrated experimentally, but so far have shown relatively poor coupling efficiency. However, the theory of these transducers has been developed and it can be shown that, with the right configuration, transducer efficiencies comparable to those obtained with lithium niobate can be achieved. Thus further development should lead to new kinds of surface-wave devices that make use of nonpiezoelectric mediums.

It would be desirable to construct acoustic surface-wave devices on a material such as silicon so that standard integrated-circuit amplification and signal processing could be carried out on the same substrate. The use of such acoustic delay lines for storage is possible, but the digital processing requires integrated circuits.

Other conceivable devices arise from interactions between surface acoustic waves and guided light.[35] There are possible new types of display devices and new methods of light modulation that make use of these interactions. Further developments in this area may give rise to an entirely new class of solid-state display devices.

It can be seen that the field of acoustic surface-wave devices is new and exciting, and makes use of many of the theoretical techniques already developed for the microwave field and experimental techniques developed for semiconductors. Practical devices are beginning to be used in radar and communication systems and studies of many new concepts are being carried out.

REFERENCES

1. Lord Rayleigh, "On waves propagated along the plane surface of an elastic solid," *Proc. London Math. Soc.*, vol. 17, pp. 4–11, 1885.

2. Rowen, J. H., "Tapped ultrasonic delay line and uses therefor," U.S. Patent 3 289 144, Nov. 29, 1966; filed Dec. 24, 1963.

3. White, R. M., and Voltmer, F. W., "Direct piezoelectric coupling to surface elastic waves," *Appl. Phys. Letters*, vol. 7, pp. 314–316, Dec. 15, 1965.

4. White, D. L., "Sonic circuits: A waveguide system for ultrasonic surface waves," presented at the IEEE Ultrasonics Symposium, Vancouver, B.C., Canada, Oct. 4–6, 1967.

5. Collins, J. H., Lakin, K. M., Quate, C. F., and Shaw, H. J., "Amplification of acoustic surface waves with adjacent semiconductor and piezoelectric crystals," *Appl. Phys. Letters*, vol. 13, pp. 314–316, Nov. 1, 1968.

6. Quate, C. F., and Thompson, R. B., "Convolution and correlation in real time with nonlinear acoustics," *Appl. Phys. Letters*, vol. 16, p. 494, 1970.

7. Luukkala, M., and Kino, G. S., "Convolution and time inversion using parametric interactions of acoustic surface waves," *Appl. Phys. Letters*, vol. 18, pp. 393–394, May 1, 1971.

8. White, R. M., "Surface elastic waves," *Proc. IEEE*, vol. 58, pp. 1238–1276, Aug. 1970.

9. Haddad, G. I., and Bahr, A. J., eds., "Special issue on microwave acoustics," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, Nov. 1969.

10. Smith, W. R., Gerard, H. M., Collins, J. H., Reeder, T. M., and Shaw, H. J., "Analysis of interdigital surface wave transducers by use of an equivalent circuit model," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, pp. 856–864, Nov. 1969.

11. Smith, W. R., Gerard, H. M., Collins, J. H., Reeder, T. M., and Shaw, H. J., "Design of surface wave delay lines with interdigital transducers," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, pp. 865–873, Nov. 1969.

12. Schulz, M. B., Matsinger, B. J., and Holland, M. G., "Temperature dependence of surface acoustic wave velocity on α quartz," *J. Appl. Phys.*, vol. 41, p. 2755, 1970.

13. Lean, E. G., and Broers, A. N., "Microwave surface acoustic delay lines," *Microwave J.*, vol. 13, p. 97, Mar. 1970.

14. Hartemann, P., and Dieulesaint, E., "Acoustic surface wave filters," *Electron. Letters*, vol. 5, pp. 657–658, Dec. 1969.

15. Hartemann, P., and Dieulesaint, E., "Intrinsic compensation of sidelobes in a dispersive acoustic delay line," *Electron. Letters*, vol. 5, pp. 219–220, May 1969.

16. Van de Vaart, H., Private communication.

17. Tancrell, R., and Holland, M. G., "Acoustic surface wave filter," *Proc. IEEE*, vol. 59, pp. 393–409, Mar. 1971.

18. Viktorov, I. A., *Rayleigh and Lamb Waves.* New York: Plenum, 1967, p. 67.

19. Tiersten, H. F., "Elastic surface waves guided by thin films," *J. Appl. Phys.*, vol. 40, p. 770, 1969.

20. Bleustein, J. L., "A new surface wave in piezoelectric materials," *Appl. Phys. Letters*, vol. 13, p. 412, 1968.

21. Gulayev, T. V., "Electroacoustic surface waves in solids," *Sov. Phys. JETP*, vol. 9, pp. 63–65, Jan. 5, 1969.

22. Matthews, H., and Van de Vaart, H., "Magnetoelastic Love waves," *Appl. Phys. Letters*, vol. 15, p. 373, 1969.

23. Tournois, P., and Lardat, C., "Love wave dispersive delay lines for wide-band pulse compression," *IEEE Trans. Sonics and Ultrasonics*, vol. SU-16, pp. 107–117, July 1969.

24. Adkins, L. R., and Hughes, A. J., "Elastic surface waves guided by thin films: Gold on fused quartz," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, pp. 904–911, Nov. 1969.

25. Van Duzer, T., "Lenses and graded films for focusing and guiding acoustic surface waves," *Proc. IEEE*, vol. 58, pp. 1230–1237, Aug. 1970.

26. White, R. M., and Voltmer, F. W., "Ultrasonic surface-wave amplification in cadmium sulfide," *Appl. Phys. Letters*, vol. 8, pp. 40–42, Jan. 15, 1966.

27. Kino, G. S., and Reeder, T. M., "A normal mode theory for the Rayleigh wave amplifier," to be published.

28. Coldren, L. A., and Kino, G. S., "Monolithic acoustic surface wave amplifier," *Appl. Phys. Letters*, pp. 317–319, Apr. 15, 1971.

29. Lean, E. G., and Tseng, C. C., "Nonlinear effects in surface acoustic waves," *J. Appl. Phys.*, vol. 41, pp. 3912–3917, 1970.

30. Svaasand, L. O., "Interaction between elastic surface waves in piezoelectric materials," *Appl. Phys. Letters*, vol. 15, pp. 300–302, Nov. 1, 1969.

31. Chao, G., "Parametric amplification of surface acoustic waves," *Appl. Phys. Letters*, vol. 16, p. 399, 1970.

32. Adkins, L. R., and Hughes, A. J., "Guided surface acoustic wave long delay lines," *Appl. Phys. Letters*, to be published.

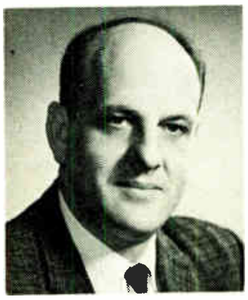
33. Lewis, M. F., and Patterson, E., "Novel helical-path surface wave delay line," *Appl. Phys. Letters*, vol. 18, pp. 143–145, Feb. 15, 1971.

34. Bond, W. L., Reeder, T., and Shaw, H. J., "Wraparound surface wave delay lines," *Electron. Letters*, vol. 7, p. 79, Feb. 11, 1971; also private communication.

35. Kuhn, L., *et al.*, "Deflection of an optical guided wave by a surface acoustic wave," *Appl. Phys. Letters*, vol. 17, p. 265, 1970.

Reprints of this article (No. X71-081) are available to readers. Please use the order form on page 8, which gives information and prices.

Gordon S. Kino (F) received the B.Sc. and M.Sc. degrees in mathematics from London University in England in 1948 and 1950 respectively. In 1955 he received the Ph.D. degree in electrical engineering from Stanford University. He joined the Mullard Radio Valve Company in Salfords, Surrey, England, in 1947, where he did research work on microwave triodes, traveling-wave tubes, and klystrons. From 1951 to 1955 he was employed as a research assistant at Stanford University, where he specialized in microwave tube research. After a short period with Bell Telephone Laboratories, in 1957 he returned to Stanford, where he is now a professor of electrical engineering. His present interests are in microwave acoustics and solid-state devices. Dr. Kino is a Fellow of the American Physical Society.



Herbert Matthews (M) received the B.A. degree from the University of Chicago in 1950, and until 1953 worked on the design and application of analog computers and hydraulic servomechanisms for the Askania Regulator Co. in Chicago. He received the M.S. and Ph.D. degrees in 1955 and 1961, respectively, from the University of Oregon. He joined the technical staff of Bell Telephone Laboratories in 1960, where he did experimental research on the elastic, microwave, and optical properties of ferrimagnets. In 1966 he joined the research staff of the Sperry Rand Research Center, Sudbury, Mass. He holds patents on several magnetoelastic devices and is currently working on elastic surface-wave physics and relevant applications. Dr. Matthews is a member of the American Physical Society, Pi Mu Epsilon, and Sigma Xi, and is serving as technical program committee chairman for the 1971 IEEE Ultrasonics Symposium.

Engineer salaries

What engineers are paid for the performance of their professional duties varies with length of experience. Graduate degree level, geographical location, and assumption of supervisory responsibility are factors in departure of salary amounts from average. There are even differences among average salaries paid to engineers in different areas of the electrical/electronics field

Just as a prudent investor continually reviews the status and prospects of the enterprises he is supporting, so the wise engineer periodically examines the nature of the company that employs him and tries to assess his potential with the organization. Often, however, it is difficult for the individual to obtain objective information about an extremely important aspect of employment—his salary.

Many engineers are unaware that, since 1953, the Engineering Manpower Commission of Engineers Joint Council has compiled information on salaries. Their most recent biennial study presents responses from 1109 establishments, including educational institutions, covering 230 064 engineering graduates. The information is organized in four reports: "Professional Income of Engineers—1970," $5.00; "Engineers' Salaries, Special Industry Report—1970," $35.00; "Salaries of Engineers in Education—1970," $2.00; and "Salaries of Engineers in Government—1970," $2.00. Any of these reports can be ordered by title from Engineers Joint Council, Department "P," 345 East 47 St., New York, N.Y. 10017. The first report, however, is available by special arrangement in single-copy sales at reduced price from IEEE directly to members. Such orders should be sent to IEEE following the instructions given in the box at the bottom of this page.

The information excerpted here from this ninth of a series represents an analysis based on years of experience, defined as the number of years since the baccalaureate degree. All charts but one show upper and lower decile (dashed lines), and upper and lower quartile (thin solid lines); all the charts show median (heavy solid line) salaries.

For those unfamiliar with charting terminology, the following definitions are furnished: The median is the point in any array of data above and below which half of the series is located (that is, the midpoint). The upper quartile is the point in any array of data above which one quarter of the series is located and below which three quarters of the series is located. The upper decile is that point above which one tenth of the series is located and below which nine tenths of the series is located.

As will be indicated for individual charts, the highest salaries received may greatly exceed the scale shown; in other words, the salary potential, though influencing the value of the upper decile, can be significantly above it, representing good opportunity for individual engineers.

The first chart, on the opposite page, shows trends in median salaries of engineering graduates over the past 17 years. Despite changes in the sources of some information, the size of the total sample and the consistency of participating agencies renders valid a comparison of the various biennial reports. It is apparent that engineers have been receiving dollar increases, on the average, over this time. The lowest increase for any experience group was of the order of 2 percent (from 1960 to 1962) and the highest was nearly 11 percent (1953–1956).

The dashed lines on this first chart indicate how the salaries of three individuals would have reacted to the combined forces of a rising economy and accumulation of experience with its concomitant rewards. Individual

Members of IEEE who wish to purchase a single copy of the 92-page report (99 charts), "Professional Income of Engineers—1970," at the special reduced price may do so by sending check or money order for $3.00, together with their name, address, and membership number, to: IEEE, 345 East 47 St., New York, N.Y. 10017. Be sure to specify the publication by name. Payment must accompany this special order.

A has 12 years of experience as of 1970 (the cutoff date of the most recent study), individual B has 17 years, and the 1970 experience of individual C totals 27 years. Between 1966 and 1970, A received a 38 percent increase in salary, B a 30 percent increase, and C (the man longest out of college) a 21 percent increase. In every case, salary refers to base salary paid by employers before deductions, and regular allowances including cost-of-living differential, if any. Unpredictable payments for overtime work, stock options, and so on, are not included.

The chart at the lower right of this page shows the weighted national averages for all engineers; the charts at the top of page 38 show the same data broken down between supervisors and nonsupervisors. The main chart and that for nonsupervisors show a slight dropoff in median salary in the later years.

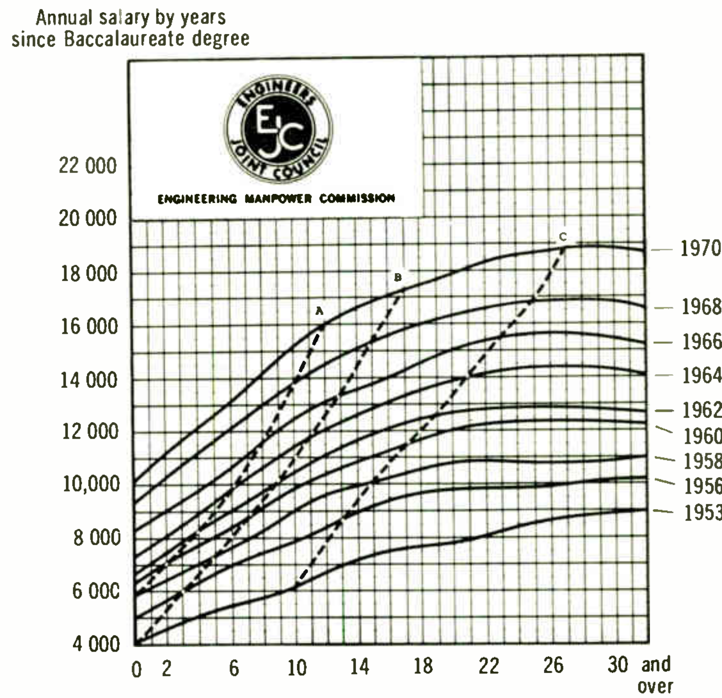
This does not mean that working engineers can expect pay cuts as they get older, however. The curves' shape is the result of the actual amount as well as the rate of pay increases in past years. The dropoff in median salaries beyond a certain point simply indicates that younger engineers tended to receive greater increases than older engineers over an extended period of time. Engineers at all levels of experience actually received increased salaries, as indicated by the dashed lines described above.

For all engineers, the median drops from a high of $18 600 to $18 000. For nonsupervisors it falls from $17 200 to $16 350. By contrast, the supervisors' median continues to rise, reaching $20 950 at 35 years. A supervisor is defined as one whose duties are primarily supervisory or managerial, whether or not they are in specifically engineering positions.

Another factor affecting these curves is the tendency for more and more engineers to shift from nonsupervisory to supervisory positions as they become more experienced. This accentuates the differential between those who advanced to higher positions and those who stayed at their previous level of work.

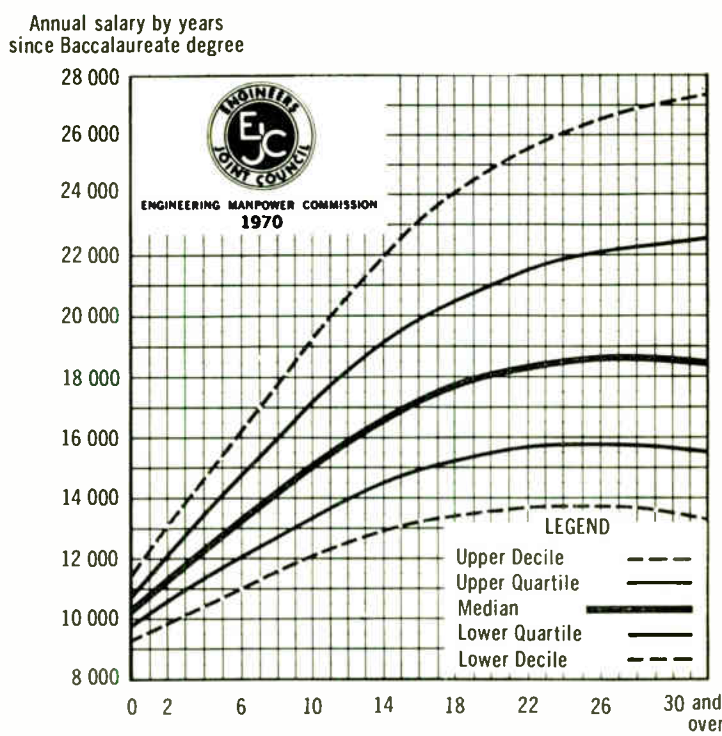
The users of the EMC studies are cautioned that there is no way of being sure whether the data furnished—and analyzed—for a particular industry group represent a typical cross section of that group. The composite charts that follow combine the salaries of supervisors and nonsupervisors and thus tend to obscure different salary patterns for the separate groups. Data in some year brackets may be based on relatively small numbers of individual salaries, thus rendering the data unrepresentative. It is believed, however, that the examples chosen for use in this article are taken from sufficiently large samples to avoid these problems.

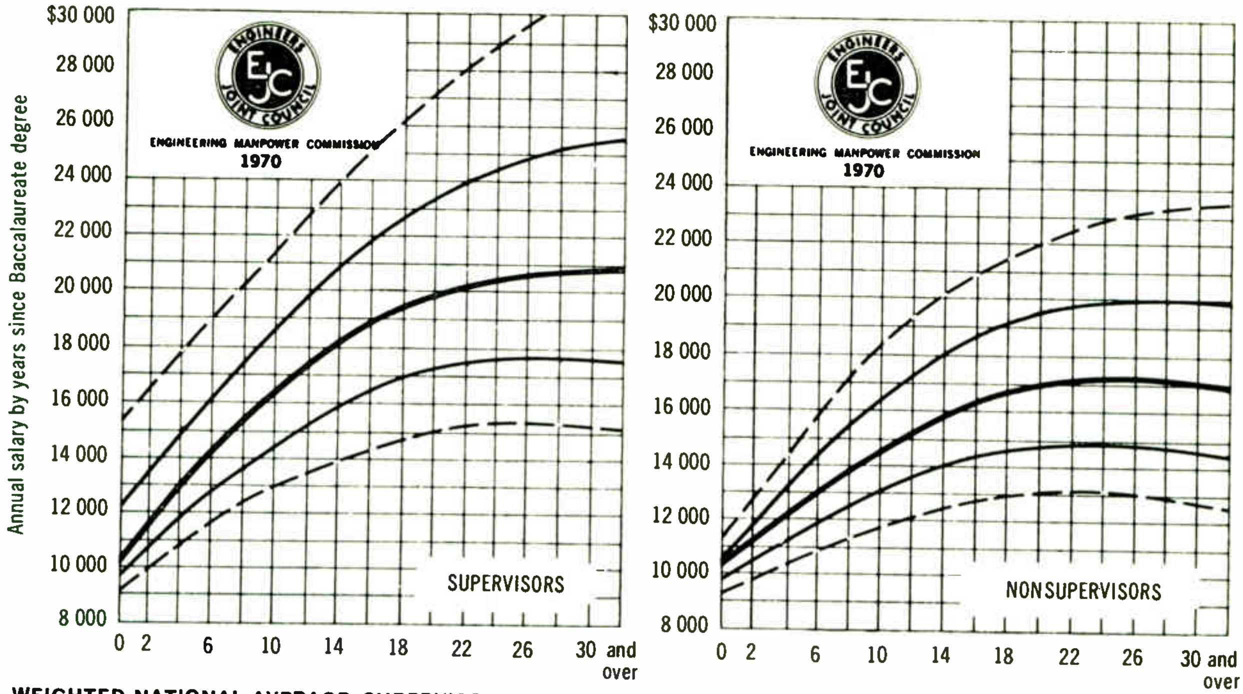
From the area described in the report as "manufacturing" is presented the chart for all engineers in the aerospace industry, 39 305 engineers covered. By comparison with the mecian for national averages, all engineers



TRENDS IN MEDIAN SALARIES OF ENGINEERING GRADUATES.

WEIGHTED NATIONAL AVERAGE, ALL ENGINEERS.



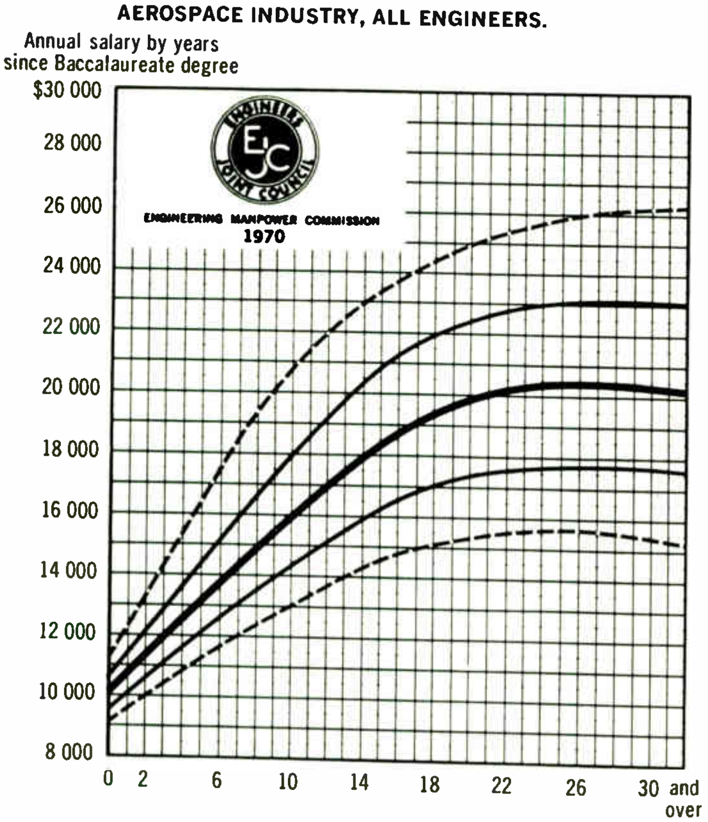


WEIGHTED NATIONAL AVERAGE, SUPERVISORS AND NONSUPERVISORS.

($18 100 at 35 years), the median here is $19 850 at 35 years. The upper decile for aerospace ($26 300) is lower than that for all engineers ($27 650) but the lower decile ($14 850) contrasts favorably with the national ($12 600).

Aerospace supervisors (total 9075 reporting) are paid better than their fellow engineers as a group with median

AEROSPACE INDUSTRY, ALL ENGINEERS.



at $20 100, upper decile $28 600, and lower decile $18 400. Nonsupervisors (total 30 230) have a median of $18 750, upper decile $23 750, and lower decile $14 350. All these salaries are at the 35-year level.

The numbers quoted have not been read from the charts but are separately printed in tables associated with them. However, the reader can scale off representative salaries for the lower-experience levels directly from the charts that have been reproduced from the EMC report. Note that the data for all charts except the first have been smoothed by fitting the raw values to a four-parameter curve calculated by the least-squares method. Since each curve is smooth individually, some of the differential between specific points may be due more to the operation of the smoothing program than to other factors.

Data developed from responses from 26 463 engineers in the electronic equipment industry make apparent the favorable economic position of that group vis-à-vis all-engineer national averages. The median of $21 300 is topped by an upper decile of $31 800 (not shown on the chart). Besides, salaries reported in excess of $39 000 are paid to a significant number of individuals with experience from 18 to 29 years. The situation for supervisors includes a median of $25 900, upper decile $36 300, and lower decile $18 450.

Engineers in the electric equipment manufacturing industry are represented by original returns of 43 327. These responses have been weighted to insure that no single company's data account for more than one half the total. The number of engineers covered is figured at 72 942. The median salary for the years 18–20 is $17 600, with the upper decile $23 400, and the lower decile $13 400. At the 35-year level the median is $17 300, upper decile $27 450, and lower decile $12 250.

The downtrend for median and lower decile in the later years—the curves flatten in the mid-twenties—is apparent even for supervisory engineers. However, these values are still higher than for the 20-year group. For them, the 18–20 median is $20 200, upper decile $25 500, and the lower decile $16 600. The 35-year median is $20 850, upper decile $31 050, and lower decile $16 150. In many cases, salaries above $39 000 are reported.

Engineers in instruments manufacturing, represented by a small sample (1396 returns), might be considered an affluent class but turn out, statistically at least, to be extremely close to the weighted national average for all engineers (no chart is shown for instrument engineers; the national-average chart is the second in this article).

The 20-year median in instruments of $17 500 compares with the national median of $17 900 and the 35-year medians are, respectively, $18 650 and $18 100. The upper deciles at 20 years are $22 800 and $24 500, whereas at 35 years, $24 950 compares unfavorably with $27 650.

Supervisory instruments engineers have a 20-year median at $19 100 and upper decile $24 250. At 35 years, the salaries are, respectively, $22 200 and $40 750. There are few numbers reported above $39 000.

In the nonmanufacturing area, 14 025 electric utilities engineers returned figures showing a median of $18 400 for 35 years, with corresponding upper decile of $30 100 and lower decile $13 150. A relatively large number (66) at the 35-year level report salaries in excess of $39 000.

Supervisory salaries for electric utilities attain a median of $20 250, upper decile $35 300, and lower decile $14 450. Nonsupervisory salaries attain a median of $15 750, upper decile $20 100, and lower decile $12 200. Charts and tables not shown in this article appear in the EMC report from which it has been excerpted.

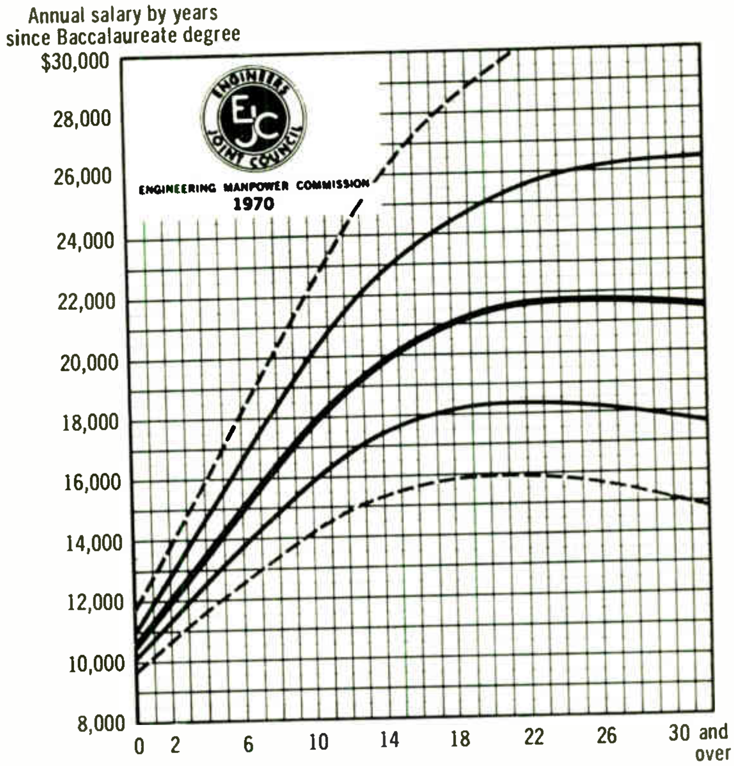
Data for 12 694 research and development engineers show a median of $20 450, upper decile $31 100 (off the chart as drawn), and lower decile $14 500. As might be expected, supervisory engineers have a good median: $25 100; the upper decile is $35 250 and the lower decile $18 200. About ten engineers in each of the groups covering 18 to 35 years' experience report salaries in excess of $39 000.

For the student or young engineer bedeviled by the desire to put away the books, get going on a job, and start a family, the charts relating the economic result of obtaining a master's or doctor's degree provide interesting guidance even if they do not solve any problems.

After 20 years, the bachelor's degree holder at a median of $17 450 can reflect that getting his master's could have put him on a curve at $19 400, whereas a doctorate could mean $20 200. At the 35-year level, the comparisons (particularly for the doctorate) are even more impressive: bachelor, $17 750; master, $19 300; and doctor, $22 550. The median curve that trends downward at 35 years for the first two is still rising for the doctor's degree.

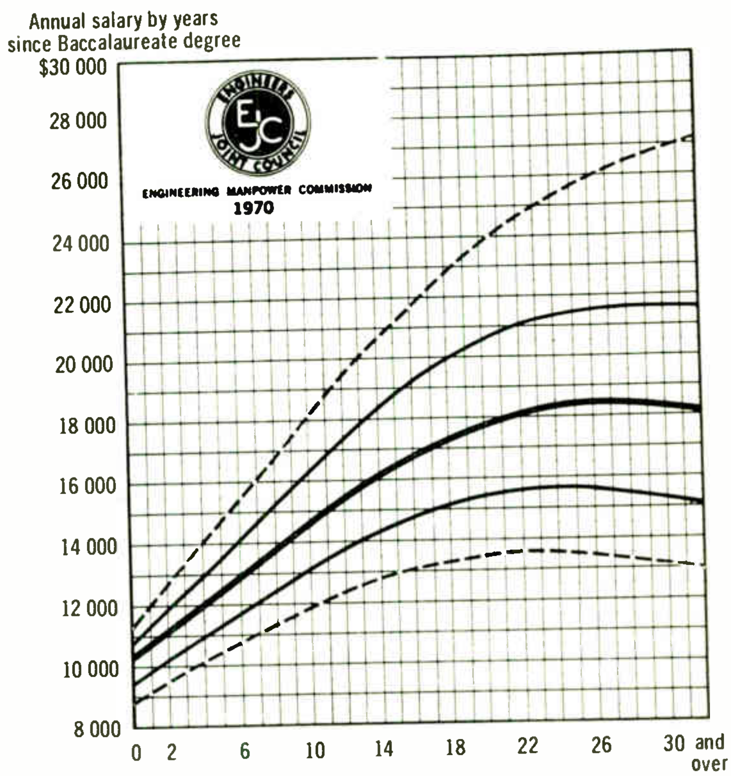
Geography often plays a subtle part in an engineer's employment. He may choose a region that he likes without regard to cost of living or salary levels and find satisfaction. If, on the other hand, he feels trapped by the economics of his location, he may discover a more congenial section or develop a good-humored resignation when he becomes aware of the economics that rule in his region.

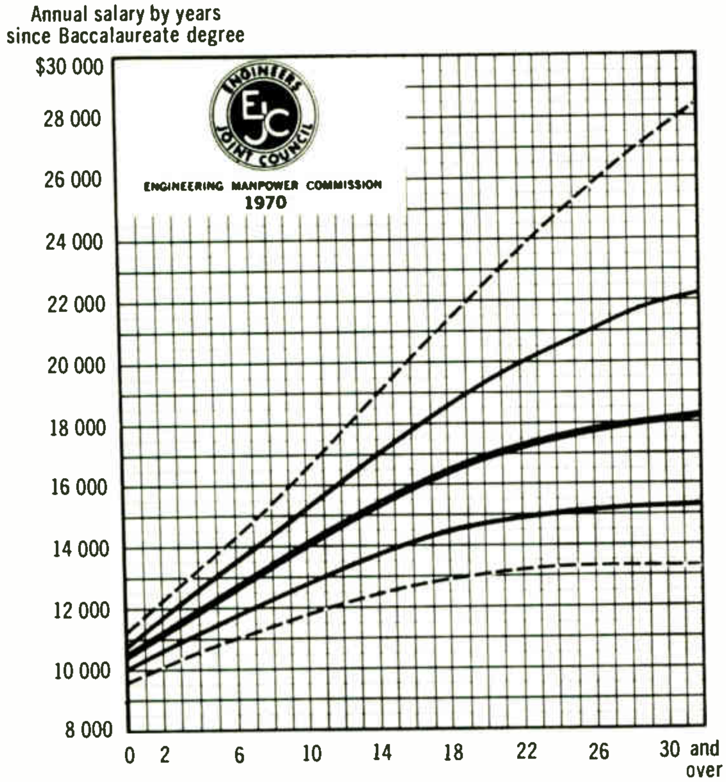
For purposes of the study, nine census regions were chosen by EMC. Information on four of these regions is excerpted here: New England, South Atlantic, Mountain.



ELECTRONIC EQUIPMENT, ALL ENGINEERS.

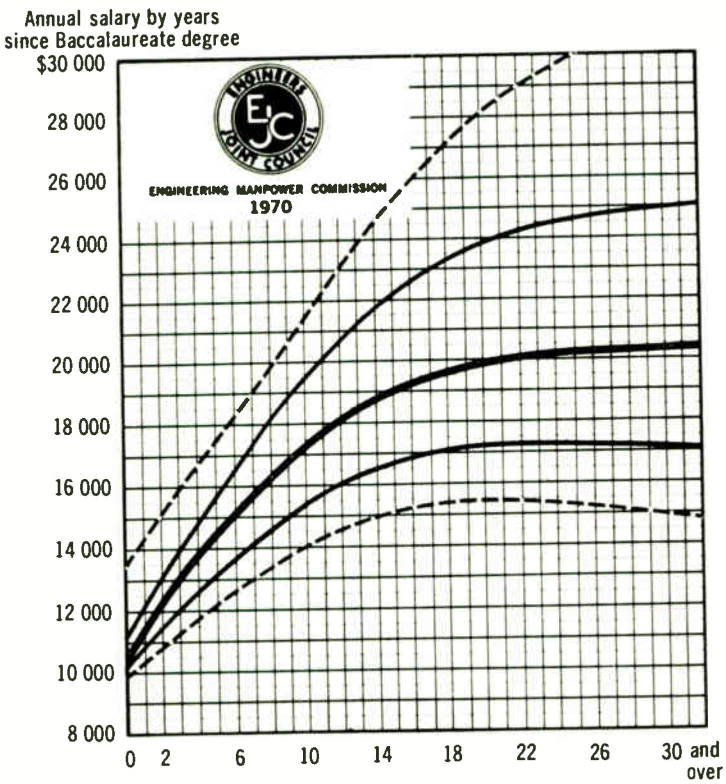
ELECTRIC EQUIPMENT, ALL ENGINEERS.





ELECTRIC UTILITIES, ALL ENGINEERS.

RESEARCH AND DEVELOPMENT, ALL ENGINEERS.



and Pacific. New England comprises Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. South Atlantic takes in Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia. The Mountain Region is Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming. In the Pacific Region, California, Oregon, and Washington are grouped with Alaska and Hawaii. The numbers of engineers responding are: New England, 9819; South Atlantic, 14 625; Mountain, 6190; Pacific, 36 690.

A simple comparison of median salaries for the 18-to-20-year experience level (where the curve is still rising) and the 35-year level in these regions is tabulated:

Region	18–20 years	35 years
New England	$17 800	$16 750
South Atlantic	18 100	18 100
Mountain	16 950	17 300
Pacific	19 550	19 550

Some of the salary differentials attributed to regional factors are also caused in part by different concentrations of industry in the various geographical areas.

It may be of interest that engineering teachers' salaries are generally comparable for all sections of the country. Industrial salaries tend to vary much more by region and by industry, whereas governmental salaries follow different patterns. Salary patterns may be influenced by the interaction of variables whose effect is impossible to separate without detailed analysis.

In publishing these curves and summary figures, IEEE might well quote EMC in stating that they attempt "to show only what prevailing salaries are, not what they ought to be. We recognize that the salary paid to any individual is, in the final analysis, affected by personal factors related to his specific duties and the way he performs them."

Engineers who have recently developed interest in studies of professional salaries may wish to review an article by Richard P. Howell of Stanford Research Institute titled "On Professional Salaries" (SPECTRUM, Feb. 1969, pp. 22–29). Using different data obtained from job applications at about 30 establishments and random samplings of faculties of more than 500 colleges, Mr. Howell identifies a number of biographical characteristics that affect salary levels.

Another publication, "Salaries of Scientists, Engineers and Technicians—A Summary of Salary Surveys," published in June 1971, is available from the Scientific Manpower Commission, 2101 Constitution Ave., N.W., Washington, D.C. 20418 for $5.00. This 88-page report gives details of starting and advanced salaries in industry, government, and educational institutions with breakdowns by field, highest degree, sex, years since first degree, age group, category of employment, work activity, type of employer, geographic area, and academic rank.

The assistance of John D. Alden, Executive Secretary, Engineering Manpower Commission, in excerpting this report from the original survey is gratefully acknowledged. The IEEE likewise acknowledges with thanks permission granted by Engineers Joint Council to reproduce copyrighted charts and figures in this article.

Alexander A. McKenzie

Long-term continental U.S. timing system via television networks

Not only is there now a 10-μs time-synchronization system available in the United States at a nominal user cost, but the system can be improved by nearly an order of magnitude and expanded to a worldwide scale

D. D. Davis, Byron E. Blair **National Bureau of Standards**
James F. Barnaba **Newark Air Force Station**

Hundreds of atomic frequency standards and precision crystal oscillators exist in remote locations throughout the continental U.S. that are synchronized through fairly complex and costly means. Today, however, an inexpensive synchronization system is available in the form of live television broadcasts by commercial networks. In precision and accuracy, the television method is comparable to the portable atomic clock and/or Loran-C, with average day-to-day differential delays less than 1μs. Based on the results of the tests presented here, the use of nearly any solid-state television receiver and a low-cost horizontal sync pulse generator can provide 10-μs synchronizations at all times. The operation of a TV line-10 timing system, including the circuitry of auxiliary equipment, is also included. This article gives about 1½ years of substantiating data for the three major commercial networks (ABC, CBS, and NBC). There is also provision for synchronization with the NBS and/or USNO Coordinated Universal Time (UTC) scales through regularly published reports.

The need for microsecond clock synchronizations* at widely separated points is becoming increasingly important to space research, defense activities, and the varied uses of private industry. Much previous work has shown the feasibility of using television signals for time comparison, especially in Europe, where the method originated and now is used quite regularly.[1–5] This article shows how some timing needs in the 10-μs region can be met throughout much of the United States by using live broadcasts originating from the New York City studios of any or all of the three commercial television networks (ABC, CBS, and NBC). The originating networks incorporate independent atomic frequency standards (rubidium) for stabilization of the horizontal sync pulses; these pulses, received at distant points, can be used to synchronize precision clocks within limitations of the distribution mediums resulting from many independent microwave links, repeater reroutes, VHF propagation, etc. The originating network signals, broadcast without auxiliary time coding, traverse varied and lengthy paths that include hundreds of microwave radio links. Such network distribution systems have given clock synchronizations within a range of 10 μs at all times for a 1½-year period. Our data were recorded at the U.S. Naval Observatory (USNO), Washington, D.C.; Newark Air Force Station (NAFS), Newark, Ohio; and the National Bureau of Standards (NBS), Boulder, Colo.

This version of television timing employs line-10 (tenth line of the odd field) in the 525-line system *M* (FCC standard for the U.S. and one of some 12 worldwide systems; see Westman[6]) as a passive transfer pulse. Almost any type of television receiver, black and white or color, is suitable for reception of signals for synchronization. Auxiliary equipment includes a line-10 synchronized pulse generator (available at a cost of about $165), a 10-MHz digital counter–printer, and a precision clock with a stability of 10^{-9} or better and having an output of 1 pps (pulse per second). This article portrays the system, including the distribution paths for the three networks, the line-10 identification circuitry, and reception results for the three labs over a period of about 1½ years. In addition, both advantages and disadvantages of the system are outlined.

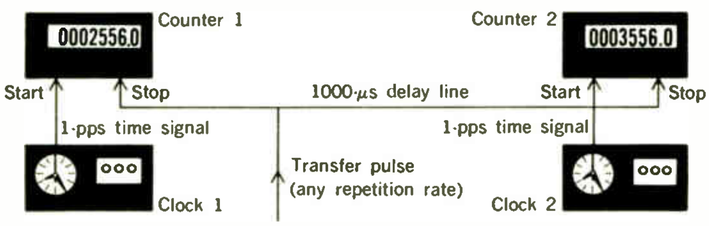
Basic concepts

For an introduction to clock comparison, consider two clocks side by side in the same laboratory, each one connected to a digital counter as shown in Fig. 1. When the 1-pps time ticks from both clocks are coincident, counters 1 and 2 will start at the same instant. Now, with a 1000-μs delay line connected between the stop inputs, the first received transfer pulse will stop counter 1 1000 μs before counter 2. The actual counter readings have no real significance; however, the *difference* in readings will be a constant 1000 μs. Conversely, a known delay would enable synchronization of clocks that may not be on time. It is possible to transfer this basic concept to precise clocks separated by several kilometers but within the service area of the same television transmitter. Once the radio propagation path has been calibrated, the television

* Synchronization is used in this article to denote simultaneity of clock readings within some frame of reference. We do not mean to imply that the method can be used to set clocks at remote locations in the absolute calibration sense.

timing system can be used to compare two or more clocks quite readily. NBS has used television horizontal sync pulses routinely since May 1968 to coordinate the clocks at stations WWV/WWVB/WWVL, Ft. Collins, Colo., with the NBS master clock at Boulder.[7] Since the horizontal sync pulses occur at 63.5-μs intervals, there is a system ambiguity over this time spacing. Such an interval corresponds to a distance of approximately 19 km (the signal travels at the velocity of light). Although the NBS radio station complex at Ft. Collins is separated by about four horizontal sync pulse periods from the Boulder Labs, the clocks agree within a small fraction of the ambiguous interval. The accuracy of such data is better than 1 μs with an rms day-to-day deviation of about 30 ns.[8]

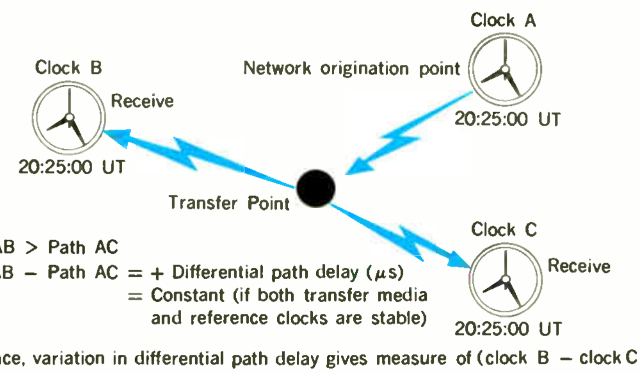
Extending the clock-comparison system one step further, we arrive at the system described in this article. Figure 2 gives the basic concept of the TV line-10 differential delay system. In our study, there is a modified television receiver, a line-10 pulse generator, a clock, and a counter at three remotely located laboratories. At the same time of day to the nearest second, counters are started at all laboratories with a 1-pps tick from their local atomic clocks. Close to this time, a horizontal sync pulse is broadcast from one of the originating television transmitters in New York City. After diverse delays through both common and separate microwave links, the sync pulse is received (live) at different times by the three laboratories and stops the appropriate counters. As in the two-clock situation in one laboratory, the difference between each pair of counter readings remains constant within the bounds of propagation delay stability of the distribution mediums and gives an accurate comparison between clocks separated by thousands of kilometers. Similarly, any laboratory can compare its clocks with NBS and USNO time scales through use of a duplicate reception system once the clock has been initially compared with a master clock and the propagation path delay has been calibrated. (Note that the clocks must be accurate to within one television picture frame or approximately 30 ms for this system to perform.)



Case 1 (Lab): Counter 2 − Counter 1 = 3556.0 μs − 2556.0 μs = 1000 μs
Case 2 (TV): Counter 2 − Counter 1 = differential path delay + clock difference (2 relative to 1)

FIGURE 1. Basic clock comparisons with delay times.

FIGURE 2. Concept of television line-10 differential measurements.



Path AB > Path AC
Path AB − Path AC = + Differential path delay (μs)
= Constant (if both transfer media and reference clocks are stable)

Hence, variation in differential path delay gives measure of (clock B − clock C)

Originating television transmission

Color television broadcasters have found that they need extremely close tolerances with regard to the phase of the color burst frequency in their televised signal.[9] To accomplish this, they have installed rubidium frequency generators at the originating New York City stations of each of the networks (ABC, CBS, and NBC). The frequency control system locks the repetition rate of the horizontal synchronization pulses, as transmitted, to the stability of the atomic reference source, and this stability can be realized by receivers at distant points. Television signals originating in New York City for national distribution are frequency-modulated, translated to microwave frequencies, and transmitted via a directional antenna to the first of a long chain of microwave repeater stations throughout the United States.

Television distribution system

Figure 3 diagrams how a television signal from an originating transmitter reaches a receiver that is thousands of kilometers away. The propagation mediums and microwave radio relay stations are the main sources of delay. This relay system consists of a chain of broadband radio links encompassing the continental United States at line-of-sight distances of some 40 to 60 km between repeaters. The estimated routing for television signals of the three networks studied in this article is shown in Fig. 4. This mapping shows that a New York City originating signal follows quite diverse paths for the three different networks in arriving at Denver, Colo.; Newark, Ohio; and Washington, D.C.

The microwave relay system carrying over 95 percent of U.S. intercity television programs is known as the TD-2 system.[10] The Bell Telephone Company developed the initial TD series in the late 1940s, and it has grown with major improvements to its present nationwide coverage of about 67 000 route-kilometers.[11] Since a detailed description of the TD-2 (vacuum-tube) system or a later modification called the TD-3 (solid-state) system is beyond the scope of this article, the reader is referred to Berger,[10] Dickieson,[12] and Sherman.[13] In brief, the system consists of hundreds of microwave repeater stations that both receive and retransmit signals in the 3.7–4.2-GHz frequency band. The initial system handled 12 RF channels, six of which were beamed forward and six backward to the appropriate repeater relay station by means of oppositely directed microwave antennas. Recent improvements have doubled this channel capacity. At each relay station the signals are processed, including amplification and delay equalization. The repeater stations are unattended, and they include standby power facilities together with automated alarm and execution systems for alerting personnel, correcting faults, and rerouting paths through special maintenance centers. At a terminating station, such as an affiliate local transmitter, the microwave signal from the closest repeater station is translated to a 70-MHz FM carrier and converted to a video signal

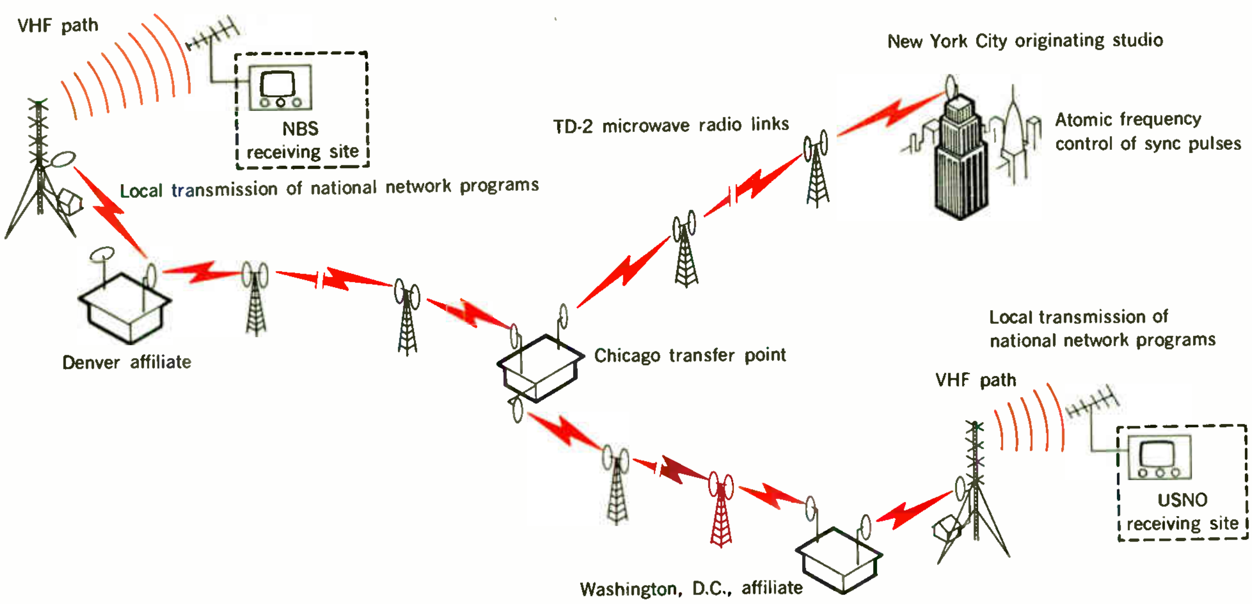
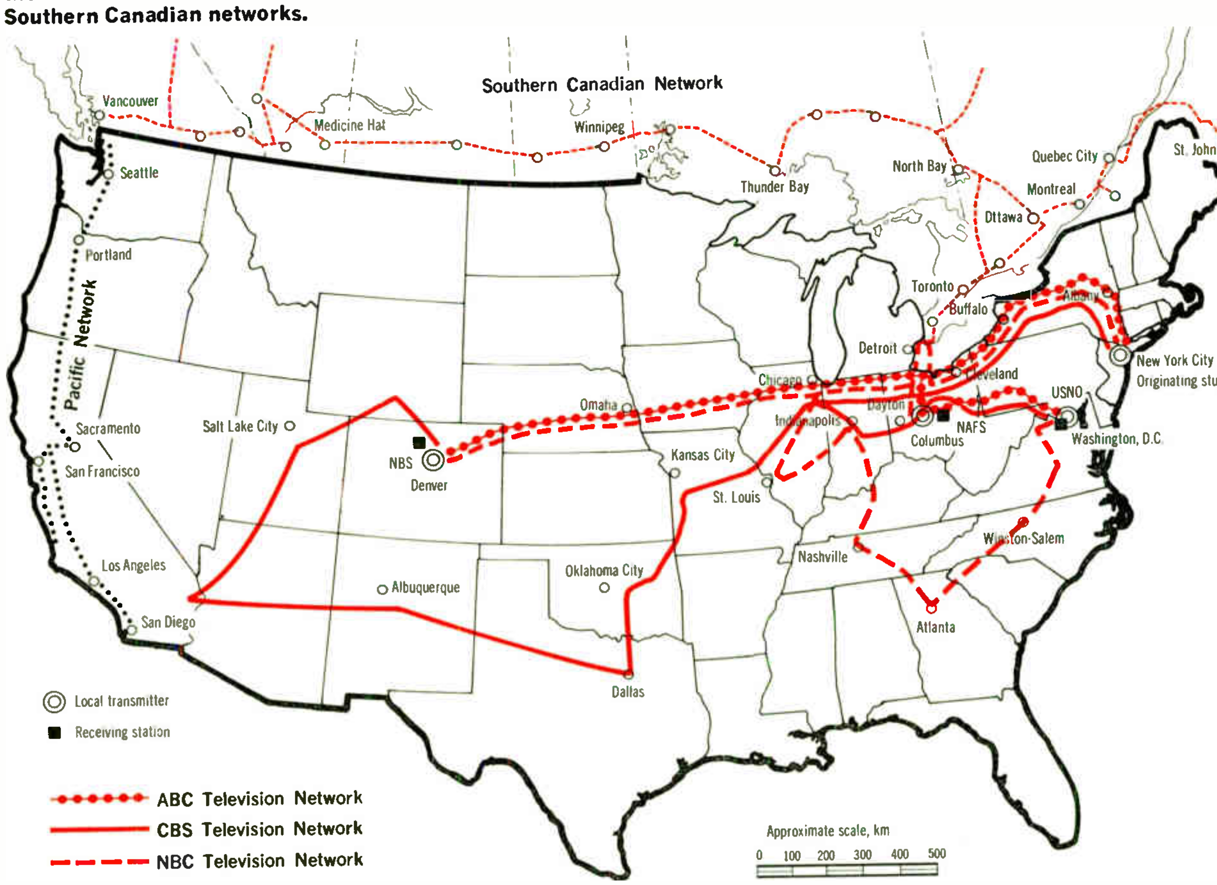


FIGURE 3. Typical routing of television signals from New York City to distant receivers. Microwave path from New York to Denver ≈ 4000–6400 km (depending on network). VHF path from Denver transmitter to NBS ≈ 30 km.

FIGURE 4. Estimated microwave radio relay routing across the continental United States for the three commercial television networks (ABC, CBS, and NBC), and the Pacific and Southern Canadian networks.



for retransmission in the VHF or UHF frequency bands (commercial television) to a local area. The reception of such local television signals at NAFS, NBS, and USNO provides the fundamental data for this article. The long-term stability shown in these data documents the possibility of disseminating accurate time via this system.

Local television reception

Figure 5 gives an overview of the television timing method and shows the equipment needed for line-10 clock synchronization at a local receiving point. The line-10 identification will be described in more detail; additional requirements for line-10 clock comparisons are:

1. All line-10 time interval measurements at different locations are made while all receiving sites are looking at the same program originating from a common source and at the same time to the nearest second. (Currently, some four hours of television programming originate from the three network stations in New York City and can be viewed simultaneously at Boulder, Colo.; Washington, D.C.; and various points in between.)

2. Since the line-10 sync pulses are ambiguous to about 33 ms, clocks at receiving points must be synchronized at least this well to ensure that the same pulse is being measured.

3. To relate remote clocks to the NBS and/or USNO time scale, it is necessary to have a measure of the differential RF delay between a receiving site and NBS or the USNO, as the case may be.

4. A highly directional television antenna is strongly recommended in areas where multipath interference can or does exist. It is believed that a 5000-μV signal level is adequate for line-10 measurements. "Ghosting," caused by secondary carrier reflection, as well as "snow" or noise in the received signals are factors to be avoided. We have experienced erroneous readings when either of these conditions existed. Quite often, realignment of the directional antenna and/or installation of a preamplifier will remedy such conditions. In addition, anomalous propagation may influence the usefulness of this television timing method in some parts of the United States. Tropospheric effects, caused by moisture and temperature inversions, called "ducting," can interfere with local television reception, particularly along the U.S. coasts. Similar interference can occur from ionospheric effects, such as sporadic *E*.

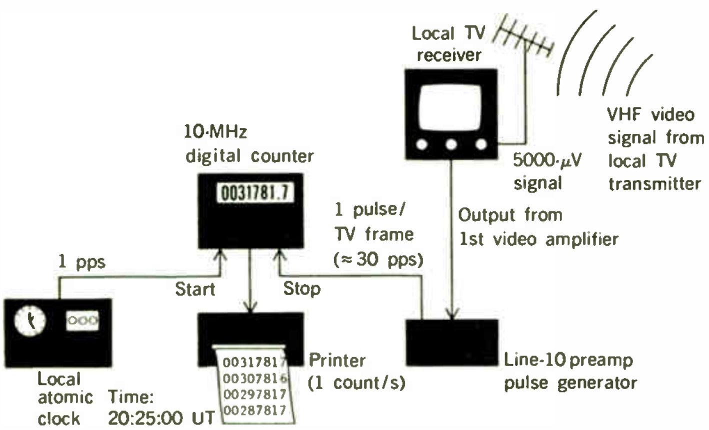


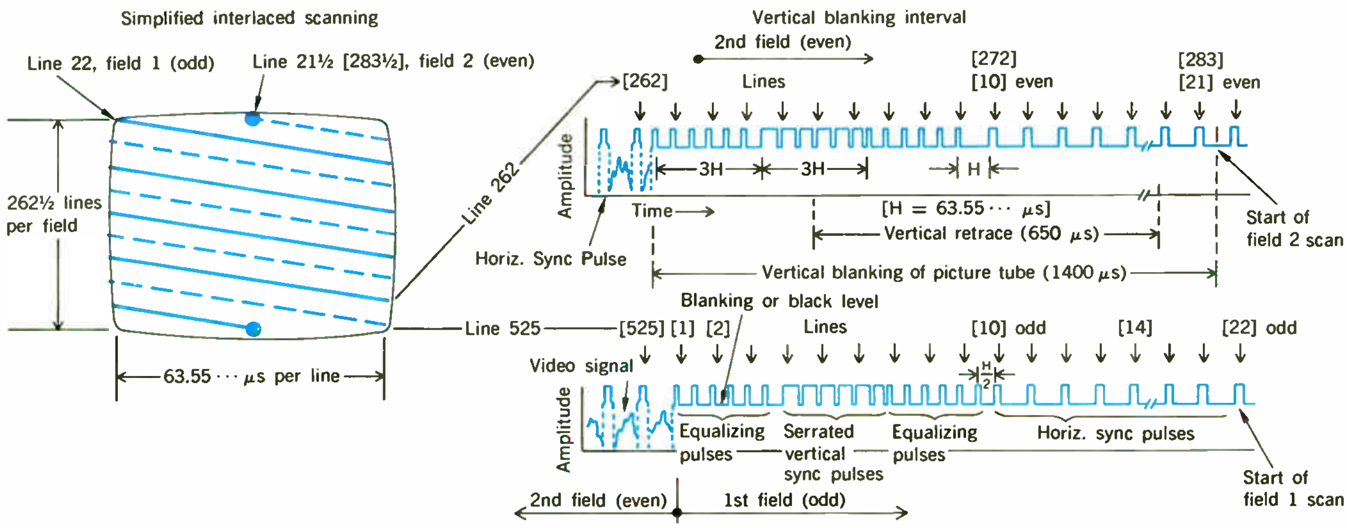
FIGURE 5. Overview of receiving laboratory instrumentation for TV line-10 synchronization.

Theory of line-10 operation

For a moment, let's consider the method used to synchronize and interlace a received television picture, as depicted in Fig. 6. A television receiver uses two sweep circuits—one to sweep the electron beam horizontally across the screen and return at a rate of 15 734.26··· Hz (63.55··· μs per line for color), and another to sweep the beam vertically down the screen and return at approximately 59.94 Hz or in 16 683.33··· μs of UTC (1970).[6] In this country, 525 lines per frame of picture is standard. However, to minimize flicker effects the frame is scanned twice in two fields with half the total number of lines in each field. Thus field 1 (or odd field) consists of 262.5 horizontal lines that scan the screen from top to bottom and return in 16 683.33··· μs of UTC (1970). On the second (even) field of display, the vertical sync pulse is displaced one-half horizontal line so that the retrace begins sooner and sandwiches or interlaces the horizontal lines of field 1. After each vertical scan of the beam tube between successive fields, the television screen is blanked for 1250 to 1400 μs. During this time, vertical retrace of the beam occurs, and signals are received in lines 1 through 9 of either odd or even fields, which contain equalizing and vertical sync pulses. Lines 10–20 also are held at the black or blanking level; however, line 10 was chosen for synchronization as it is the first horizontal sync pulse after the equalizing and vertical sync pulses and therefore is easy to identify with simple logic circuits.

The line-10 identification circuitry was designed to generate one pulse for each frame (525 lines) of video. To accomplish this, it is necessary to distinguish between line 10 of the odd and even fields. Our synchronization circuitry identifies the trailing edge of line 10 in the odd

FIGURE 6. Interlaced scanning of television picture and vertical blanking interval format.



field and, as indicated in Fig. 6, this can be easily accomplished through the time relationship of line 10 to the equalizing pulses of the successive fields.

Line-10 identification (ID) is effected by integrated circuitry such as is shown in Fig. 7. The major components consist of a combination preamp–sync stripper, an odd-field line-10 pulse generator, and a power supply. The input to the preamp is a composite video signal taken from the first video amplifier emitter follower of a typical solid-state television receiver. (Nearly any television set can be used provided the signal level, impedance, and video polarity are compatible; older tube-type sets require additional impedance-matching circuitry.) The video signal at this stage is positive (negative-going sync) with a peak-to-peak amplitude of 0.6 to 1.0 volt. The operation of the line-10 ID circuitry is described in some detail in the appendix. Several options of this unit are available locally; a completely assembled and tested unit (preamp, line-10 pulse generator, and power supply) costs about $165 (the cost of parts alone is about $50).

Data acquisition

The line-10 data are obtained from the instrumentation that is shown in Fig. 5. Only one reading is recorded for a given network and time. Actually, we record readings for the 39 seconds after the specified starting time at one-second intervals to average the variation in the 0.1-μs column. A typical printout is shown in Fig. 8. Measurements of NBC, for instance, are commenced at 20:25:00 UT and each second thereafter until 20:25:39 UT. The 40 readings are averaged to minimize error in the 0.1-μs column. In this particular example, the reading for 20:25:00 would be reported as 32 440.3 μs. Note that for each second the count increases by 1000 μs until the modular frame interval of 33 366.66$\cdots$ μs is exceeded. This results from the relationship of the television frame rate relative to 1 pps. That is, 30 vertical frames occur for 1.001 000$\cdots$ seconds of UTC (1970). The relationship is

$$\text{Period of 1 line} = \frac{1}{\text{Horiz. scanning freq.}}$$

$$= \frac{1}{\dfrac{2}{455}\left[\dfrac{63}{88} \times 5 \times 10^6\right]} = 63.55\cdots\ \mu\text{s}$$

with the term in brackets equal to the chrominance subcarrier frequency; and

Period of 30 frames (525 lines/frame)

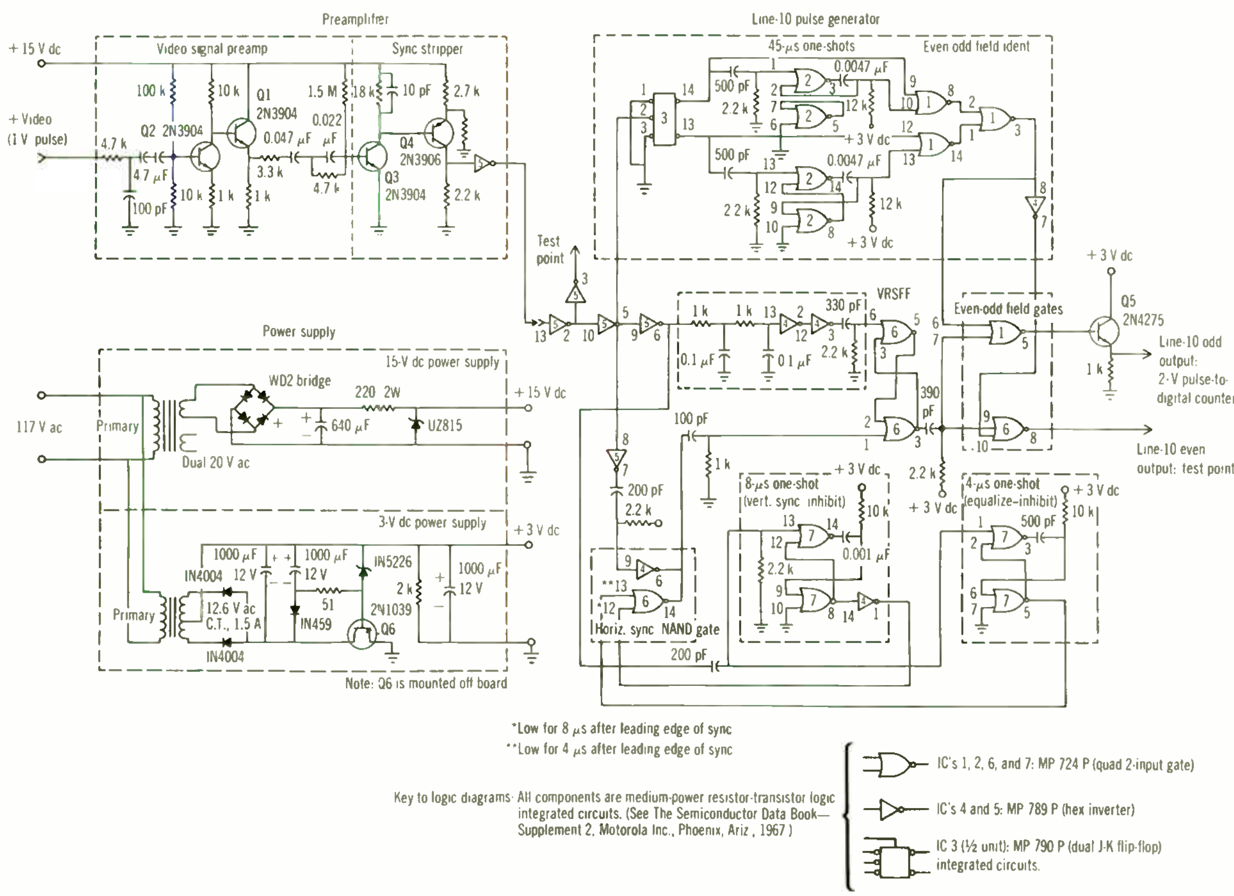
$$= \frac{1}{\left[\dfrac{2}{455} \times \dfrac{63}{88} \times 5 \times 10^6\right]} \times 525 \times 30$$

$$= 1.001\ 000\cdots \text{ seconds of UTC (1970)}$$

In the printout of Fig. 8, the set B readings can be converted to an extension of set A by adding one modular frame interval to each reading.

To use TV line-10 synchronization, it is necessary to resolve ambiguity to less than the period or module of

FIGURE 7. Circuitry of line-10 identification using integrated logic components.



one television frame. In our present studies, a maximum differential delay of about 24 800 μs occurs for the CBS network if one monitor is in New York City and the other in Boulder. Assume clocks at Boulder and New York City are on time and that both monitors start their respective counters simultaneously. The next line-10 sync pulse (corresponding with 1 pps) that originates from the New York City transmitter will stop the New York City counter. About 24 800 μs later the same sync pulse would stop the NBS counter unless a pulse was in transit at the time both counters were started. In such a case, the NBS counter would be stopped before the New York City counter. To use such data from the two counters, one must increase the NBS value by the period of one television frame to compensate for path delay. Note the following example:

	TV Reading	
NBS	2 748.2 μs	2 748.2 μs
NYC	11 314.9 μs	+33 366.7 μs
	(NBS plus corr.)	36 114.9 μs
	minus N.Y.C. to NBS delay	−11 314.9 μs
Differential delay ± clock error =		24 800.0 μs

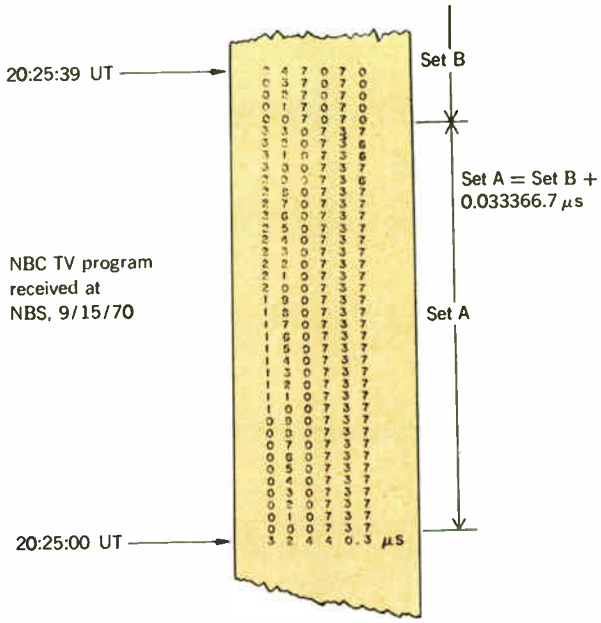


FIGURE 8. Sample printout of TV line-10 (NBS) counter readings in terms of clock 8.

Results of line-10 synchronization

During the past 15 to 18 months, television timing measurements were made via the three networks at NAFS, NBS, and USNO as outlined in Table I. Note that two measurements are made for each network, both before and after the half-hourly station breaks. This is done to study the effects of national-to-local switching and to determine if a consistent time interval elapses when the national network program resumes. In many cases the time of return is many microseconds removed from the computed counter reading based upon the frame modular count for the 6-minute interval. This could occur from minor rerouting of signal path, replacing sync generators, sync generators randomly jumping, or other causes that are unknown at this time. In any event, if the change is seen simultaneously at all monitoring stations, it will not affect the differential measurement or the actual clock comparison. From day to day, the 6-minute differential measurements can vary by several tenths of a microsecond. Some results of differential timing measurements at the three laboratories have been plotted in Fig. 9 for June 1969–December 1970. Variations in the differential delays result from differences between the controlling clocks as well as contributions from the measurement system. During this time there were seven apparent network reroutes not common to all receiving sites. The data reflect deviations from nominal; where reroutes occur, adjustments were made. Because the chances for simultaneous rerouting of each of the three networks are remote, nominal delay times can be adjusted based on the stability of the other two networks.

The NBS–NAFS data (Fig. 9A) show good agreement for the three network paths. From September 1969 to December 1970, these data fall within a maximum range of 6 μs. This variation is believed to reflect mainly the differences between the two clocks at these laboratories. The NBC data for the NBS to NAFS path show no major network rerouting and, during 1970, they fall within a maximum deviation range of 6 μs. Typically, the monthly standard deviation about the mean nominal differential delay is about 0.5 μs. Through averaging of the three network results and eliminating extraneous values, we feel that standard deviations of 0.2 μs should be possible for a monthly set of data.

The NBS–USNO data (Fig. 9B) similarly agree for the

I. Characteristics of television receiving sites

		NAFS, Ohio				NBS, Colorado				USNO, Washington, D.C.			
Network	Measurement Time, UT*	Channel	Frequency, MHz	VHF Path Dist., km	Trans. Bearing, degrees	Channel	Frequency, MHz	VHF Path Dist., km	Trans. Bearing, degrees	Channel	Frequency, MHz	VHF Path Dist., km	Trans. Bearing, degrees
NBC	20:25:00 20:31:00	4	66–72	47.5	270.9	4	66–72	29.5	175.6	4	66–72	2.6	326.3
CBS	20:26:00 20:32:00	10	192–198	47.4	263.5	7	174–180	29.6	175.8	9	186–192	2.9	305.5
ABC	20:27:00 20:33:00	6	82–88	46.2	249.7	9	186–192	29.6	175.6	7	174–180	3.5	339.4

*One hour earlier during Daylight Saving Time in the summer.

DAILY TELEVISION TIME TRANSFER MEASUREMENTS**

Listed below are the daily readings (for the three major U. S. networks) of the difference between UTC(NBS) and the trailing edge of the next line 10 (odd) horizontal synchronization pulse as received in Boulder, Colorado, at the times specified. Note: These pulses have been observed for a sufficient period of time to merit their consideration as a source of time transfer due to their low scatter. This schedule will be followed in succeeding months.

UTC(NBS) - Line 10 (odd) (in microseconds)

Sept.	NBC		CBS		ABC	
1970	19:25:00*	19:31:00*	19:26:00*	19:32:00*	19:27:00*	19:33:00*
1	04324.4	30659.2	05569.6	31902.9	09152.5	02119.7
2	18057.4	11023.7	19263.3	12230.0	10421.1	15844.9
3	31787.1	03844.2	32958.1	25924.7	03236.5	29569.8
4	12152.3	05120.2	13285.7	06252.3	16961.9	09928.7
5	-	-	-	-	-	-
6	-	-	-	-	-	-
7	-	-	-	-	-	-
8	21231.0	14198.6	01328.6	27661.7	05130.3	31463.8
9	01598.0	27930.5	15022.7	07989.3	18855.6	11822.5
10	15328.0	08296.0	29527.1	22493.8	32581.3	08644.3
11	-	-	-	-	-	-
12	-	-	-	-	-	-
13	-	-	-	-	-	-
14	-	08970.6	05859.2	32193.6	20751.4	13718.5
15	32440.2	25406.6	19554.7	12521.3	01111.1	27444.6
16	24844.8	17621.1	33249.4	26216.1	14837.1	07803.8
17	05019.6	31352.7	13577.8	06544.5	28563.3	21530.1
18	07311.0	00279.3	27273.1	20239.9	20769.8	01890.7
19	-	-	-	-	-	-
20	-	-	-	-	-	-
21	15513.3	08479.8	01625.4	27958.7	16736.3	09703.2
22	29246.0	22212.6	15320.6	08287.2	30462.9	23429.8
23	09613.3	02579.6	29015.7	21982.4	10823.1	30800.9
24	23346.2	16312.9	09344.5	02311.2	24549.9	17516.7
25	03711.1	30042.9	23039.9	16006.5	04910.4	31243.8
26	-	-	-	-	-	-
27	-	-	-	-	-	-
28	16583.7	09615.5	30758.5	32050.7	12724.1	05691.0
29	30380.3	23348.4	09253.9	02220.6	26451.3	19417.8
30	-	-	-	-	-	-

The above data is being published as a service to users of television timing systems. It in no way implies a guarantee of results from using such a timing system, nor does it signify any control of the emission times of synchronization pulses by the NBS.

*Universal Time

**See Section 8 of Bulletin No. 151 - June 1970.

three networks. A graph of corrected Loran-C data (Dana) and USNO portable clock measurements has been superimposed (relatively) on the ABC data and quite close correspondence exists. For the last half of 1969, the Loran-C data and portable-clock check deviate about 3 μs from the TV line-10 curve. However, during 1970, the variations are about 1 μs. The reason for the 1969 variation is unknown except for possible equipment difficulties. (The Loran-C data consisted of signals received from Dana, Ind., at NBS in Boulder, Colo., and corrected for Cape Fear, N.C., [master] variations as reported by the USNO.[14]) The statistics of the relationships between remote time synchronizations via portable clocks, Loran-C, and television broadcasts are published elsewhere.[15]

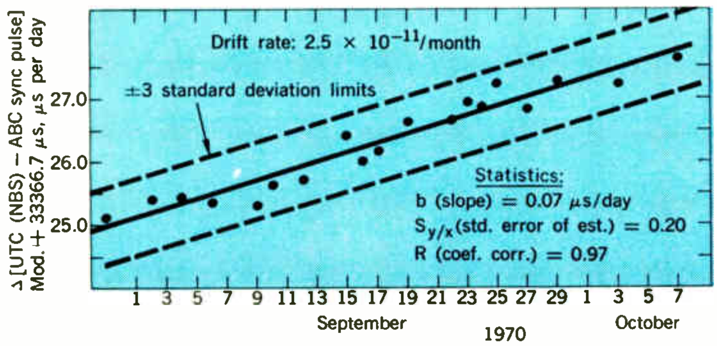
The NBC and CBS data for the NBS–USNO path show changes of 4 to 6 μs that do not appear real. (From December 1969 to January 1970, the NBC differential delay gains about 6 μs; from July to August 1970, the CBS data show a gain near 4 μs.) Without any adjustments, however, the data, with few exceptions, fall within a maximum excursion range of 10 μs.

Relation of TV line-10 measurements to NBS and USNO time scales

NBS issues to users on the basis of need the line-10 daily measurements in terms of UTC (NBS) in the monthly NBS Time and Frequency Services Bulletin.[16] A sample data sheet of NBS television measurements is shown for September 1970 in the box on page 47. The USNO distributes on the basis of need line-10 data in terms of UTC (USNO) in their weekly Series 4 time services bulletin.[14] These bulletins will publish future changes, modifications to the television system, or other factors that might affect a user in the field. Because of the periodic relationship of the television time interval with UTC, one usually can compute the value of a missing counter reading. In fact, the ABC data at Boulder show consistent continuity in counter readings between days for weeks at a time. For unknown reasons, the CBS and NBC data do not show such continuity over weekends; however, within respective weeks they are consistent. This may be due to a change of sync generators at the New York City originating stations.

Data such as shown in the sample box also enable one to study the stability of the network rubidium standards. The counter readings for UTC (NBS) minus line-10

FIGURE 10. Daily divergence between the UTC (NBS) clock and UTC (ABC) Rb clock in New York City via TV line-10 measurements.



(ABC) were adjusted by the daily modular frame rate to obtain the difference between the computed and actual readings. This difference is the daily divergence in microseconds between the NBS and ABC clocks. A plot of such results is given in Fig. 10 for the month of September 1970. This graph indicates that the rubidium standard at New York City (ABC) was drifting in terms of the UTC (NBS) clock at a rate of about 2.6 parts in 10^{11} per month with a standard error of estimate about the least square line fitting the data equal to about 0.2 μs. Rubidium standards, controlling television sync pulses over the microwave radio links, show nearly equivalent stability to laboratory specifications. This illustrates an additional potential of the TV line-10 system. With a known atomic frequency reference at the New York City originating studios and with the cooperation of the network distribution systems, counter readings in terms of UTC (NBS) or UTC (USNO) could be predicted, say a week in advance, within tenths of a microsecond.

Advantages and limitations

The advantages of TV line-10 synchronizations can be listed as follows:

1. Initial equipment is low in cost.
2. Measurements are simple.
3. The cost per clock synchronization is low compared with portable clock measurements.
4. Transmission of data has no effect on television network programs or viewing reception and is without cost to user.
5. Simultaneous clock comparisons can be made without error of a transfer portable clock.
6. The system shows a 10-μs reliability with the potential of a few microseconds.
7. Modular intervals are potentially predictable.
8. Three television networks with atomic clock references provide redundancy and enable cross synchronizations.
9. Simultaneous viewing of one television broadcast in the immediate area of several receivers can enable microsecond synchronization among clocks at these stations without regard to national "live" programming.

There are also limitations:

1. Microwave paths can be interrupted without notice.
2. There is limited simultaneous viewing time of nationwide network programs.
3. Present network distribution systems will not allow tie-in with West Coast transmission lines. However, the Los Angeles studios originate live programming several hours a day for the West Coast area; if line-10 measurements were made in terms of an NBS or USNO synchronized cesium clock and published, microsecond comparisons should be possible for this part of the United States also.
4. System will not work with tape delays—program must be "live" for all simultaneous measurements referred to the NBS and USNO television data.
5. Measurements currently are not made on weekends.
6. System ambiguity is 33 ms.
7. Propagation anomalies may limit the system's usefulness in some areas of the United States.

Conclusions

A 10-μs synchronization system now exists through the use of commercial television networks and inexpensive laboratory equipment. The television system shows potential for nearly an order-of-magnitude improvement with the cooperation of microwave radio equipment lessees (notice of reroute changes); redundant cesium frequency standards controlling originating broadcasts with known reference to UTC (NBS) or UTC (USNO); phase-continuous operation of such originating broadcast equipment as sync generators over weekends with backup battery power supplies; and improved digital counters at receiving laboratories with ns resolution. (See Parcelier.[5])

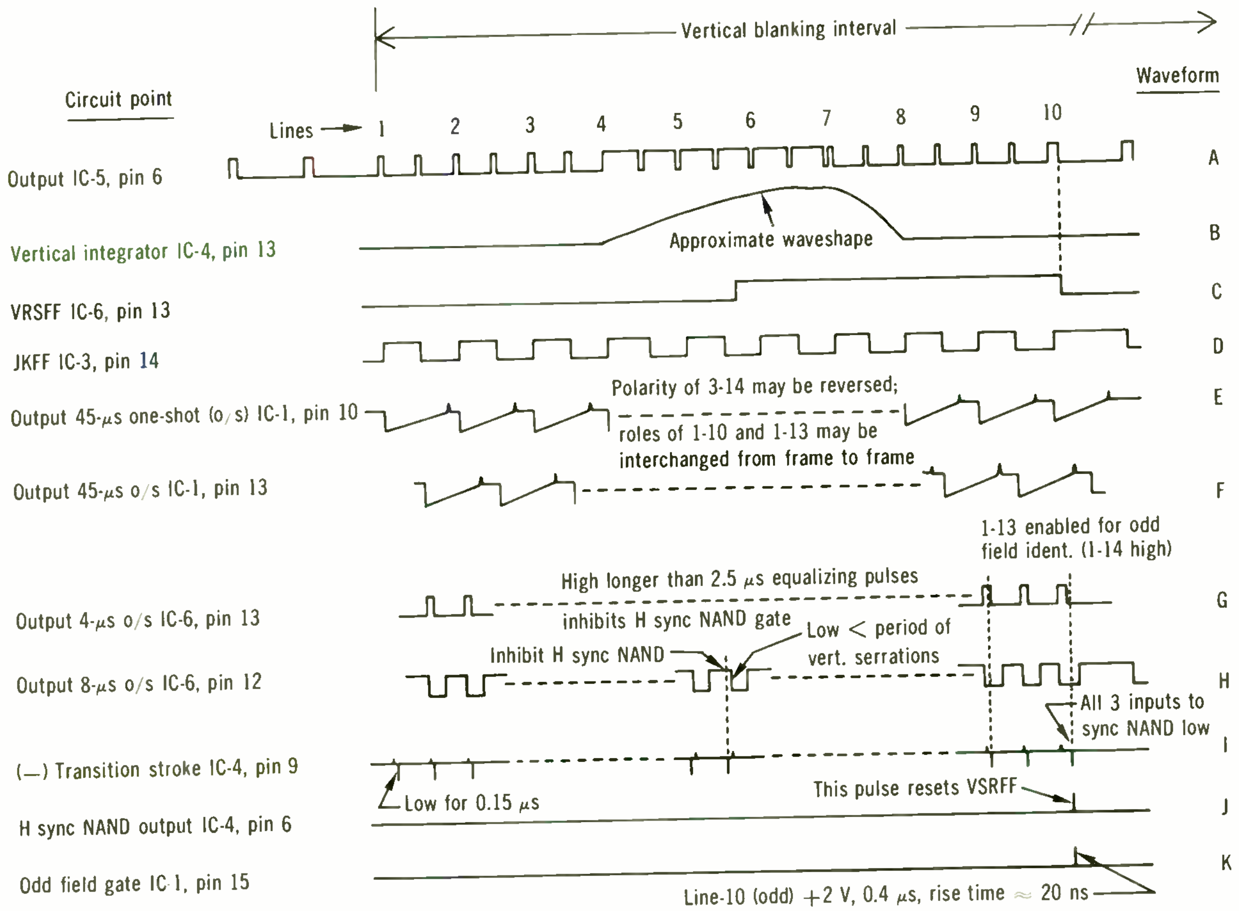
Additional and possibly more consistent data could be obtained if the line-10 system were automated to print out measurements on a daily basis. We encourage laboratories maintaining precise time, especially in southern Canada, southeastern United States, or any other intermediate points, to compare their clocks with the NBS or USNO time scales through TV line-10 synchronizations. (Since March 1, 1971, the USNO time service substation near Miami, Fla., has made TV line-10 measurements with good results.) The system conservatively covers about 70 percent of the continental United States, and it is hoped that some "live" television coverage can be extended to the West Coast distribution networks. The method is usable, however, for microsecond synchronization by simultaneous viewing of a common television broadcast; it should not be overlooked in coordinating many precise time centers, ships at sea, etc., within the broadcast field of a centrally located transmitter, such as on the West Coast or in Hawaii or Alaska.

As one looks to the future, the satellite-television links proposed for global communications show great potential for microsecond synchronization between clocks in different countries, provided basic differences in television system standards can be resolved.

Appendix. Details of TV line-10 pulse generator

As noted in this article, the input to the line-10 preamp is the positive video signal (negative-going sync) from the television set with a peak-to-peak amplitude of 0.6 to 1.0 volt. Refer to Fig. 7 for the following description of operation. The composite sync is amplified and inverted; 4 to 7 volts of composite video, with positive-going sync, is applied to the two-stage sync stripper to clean up the signal not reaching sync level. The output of the sync stripper at IC-5 (integrated circuit 5), pin 1 [sync at RTL (resistor-transistor logic) level], is connected to the input of the odd-field line-10 pulse generator. The output of IC-5, pin 6, is the composite sync waveform *A* shown in Fig. 11. The serrated vertical sync pulse, integrated by the vertical integrator, sets the VR-SFF (vertical reset–set flip-flop) at approximately the fourth serration (waveforms *B* and *C*, Fig. 11). The leading edge of every pulse will trigger the 4- and 8-μs one-shots and also toggle the dual JK flip-flop (IC-3) in the even–odd field identification circuitry. The waveform output of IC-3, pin 14, is shown as *D* in Fig. 11. If the time between leading edges is less than 45 μs, the output of IC-1, pin 3, will be low, enabling gate IC-1, pin 6. (See waveforms *E* and *F*, Fig. 11.)

FIGURE 11. Line-10 identification waveshapes at various parts of the circuit of Fig. 7.



The trailing edge of every pulse activates the horizontal sync NAND gate at IC-4, pin 9. If the time between leading and trailing edges is more than 4 μs but less than 8 μs, the horizontal sync NAND gate will generate a positive pulse that resets the VR-SFF (waveform *J*). This reset will drive IC-1, pin 7, and IC-6, pin 9, low for about 0.4 μs, resulting in a $+0.4$-μs output pulse for either L10 (line 10) odd or even, depending on the state of the even–odd identification. The output for L10 odd is therefore a $+0.4$-μs pulse coincident with the trailing edge of the L10 odd pulse (waveform *K*). The amplitude of this pulse is about 2 volts, with a rise time of $\approx$20 ns, and is sufficient to drive a time interval counter.

Component tolerances are such that no adjustments should be required after final assembly. The 45-μs one-shots should have a period between 40 and 55 μs. The range of the 8-μs one-shot period should be between 6 and 12 μs and that of the 4-μs one-shots between 3.7 and 4.3 μs.

The authors wish to thank George Kamas and James Barnes for substantive advice and helpful consultation in the preparation of this article. Wallace Hamilton and Glynetta Perrymore compiled the data and provided the data analysis. The U.S. Naval Observatory and the three television networks (ABC, CBS, and NBC) are acknowledged for their valuable help and cooperation in these tests.

REFERENCES

1. Tolman, J., Ptáček, V., Souček, A., and Stecher, R., "Microsecond clock comparison by means of TV synchronizing pulses," *IEEE Trans. Instrumentation and Measurement*, vol. IM-16, pp. 247–254, Sept. 1967.

2. Souček, A., "Travel time stability on TV relay links," *Radio Television*, no. 5, pp. 29–31, Sept. 1969.

3. Allan, D. W., Leschiutta, S., and Rovera, G., "TV frame pulses used for precision time synchronization and their noise distribution," *Alta Freq.*, vol. 34 (English issue no. 2), p. 180E, Mar. 1970.

4. Davis, D. D., Jespersen, J. L., and Kamas, G., "The use of television signals for time and frequency dissemination," *Proc. IEEE (Letters)*, vol. 58, pp. 931–933, June 1970.

5. Parcelier, P., "Time synchronization by television," *IEEE Trans. Instrumentation and Measurement*, vol. IM-19, pp. 233–238, Nov. 1970.

6. Westman, H. P., (ed.), *Reference Data for Radio Engineers*, 5th ed. Indianapolis, Ind.: Howard W. Sams & Co., 1969, pp. 28-9–28-17, 28-31.

7. Milton, J., "Standard time and frequency: its generation, control, and dissemination from the National Bureau of Standards time and frequency division," Tech. Note 379, U.S. National Bureau of Standards, Aug 1969, pp. 21–25.

8. Allan, D. W., Private communication, Feb. 1971.

9. Schmid, H., "Synchronization of remote program sources for color TV broadcasting," *J. Soc. Motion Picture Television Engrs.*, vol. 78, pp. 619–620, Aug. 1969.

10. Berger, U. S., "TD-2: eighteen years and still growing," *Bell Lab. Record*, vol. 46, pp. 210–216, July/Aug. 1968.

11. Latter, R. F., "Microwave radio and coaxial cable facilities in the long distance telephone network," *Telecommunications*, vol. 3, pp. 11–17, Feb. 1969.

12. Dickieson, A. C., "The TD story: from research to field trial; vacuum tubes and systems engineering; changing for the future," *Bell Lab. Record*, vol. 45, pp. 282–289, Oct. 1967; pp. 324–331, Nov. 1967; pp. 356–363, Dec. 1967.

13. Sherman, R. E., "The TD-3 microwave radio system," *Bell Lab. Record*, vol. 47, pp. 286–291, Oct. 1969.

14. "USNO daily phase values—Series 4," Time Service Division, U.S. Naval Observatory, Washington, D.C.

15. Allan, D. W., Blair, B. E., Davis, D. D., and Machlan, H. E., "Precision and accuracy of remote synchronization via portable clocks, Loran-C, and network television broadcasts," *Proc. 25th Ann. Symp. on Frequency Control*, Electronics Industries Association, Washington, D.C. (in press).

16. "NBS Time and Frequency Services Bulletin," Frequency–Time Broadcast Services Section, Time and Frequency Division, NBS, Boulder, Colo.

Reprints of this article (No. X71-082) are available to readers. Please use the order form on page 8, which gives information and prices.

D. D. Davis attended the University of Oklahoma, Oklahoma City University, and Oklahoma State University, and has taught in many areas at the FAA Academy in Oklahoma City. His work there included airways navigation theory and the computer-aided air traffic control system. His responsibilities included development of course material for the Air Navigation Facilities Training Branch. Since 1968, Mr. Davis has been employed in the National Bureau of Standard's Time and Frequency Division. His work at this agency has included studies of aircraft collision avoidance systems, Loran-C, and television timing. Mr. Davis has written several papers on various aspects of TV time and frequency dissemination.

Byron E. Blair (SM) received the B.S. degree in physics from Wheaton College, Wheaton, Ill., in 1942. He is at present a physicist at the National Bureau of Standards, Time and Frequency Division, where he is engaged in research on precise frequency measurement and phase stability of radio signals. Before his transfer to NBS in 1960, Mr. Blair was associated with the U.S. Bureau of Mines, College Park, Md., where he evaluated vibration instrumentation and studied varied aspects of rock mechanics. During World War II, he was on active duty as an officer in the U.S. Navy and is now a retired lieutenant commander in the U.S. Naval Reserve. Mr. Blair has contributed several papers to "Radio Science" and is coeditor of NBS Special Publication 300, "Precision Measurement and Calibration—Frequency and Time," volume 5. In August 1969, he served as a National Academy of Sciences–National Research Council delegate to the 16th general assembly of URSI in Ottawa, Canada. He is a member of the CCIR and the Scientific Research Society of America.

James F. Barnaba received the B.S.M.E. degree from Ohio University in 1958. After serving three years in the U.S. Air Force as an air transportation officer, he joined the engineering staff at Newark Air Force Station, Newark, Ohio. During the period from 1961 to 1967, he provided engineering support for the repair and calibration of Titan missile inertial guidance systems and F-4C inertial navigation systems. In 1967, Mr. Barnaba transferred to metrology engineering (electrical branch) at Newark AFS and became involved in precise time and time interval work. His present responsibilities include time transfer via atomic clocks, television pulses, and satellite transmissions.

Automatic word recognition

The continuing effort to bridge the man–machine communication gap may be further assisted by the development of an automatic word recognizer that uses an adaptive memory technique to adjust to each speaker

G. L. Clapper *International Business Machines Corporation*

Separately spoken individual words can be automatically recognized using a two-dimensional pattern of spectral density versus a nonlinear time base. The pattern for a given word differs from person to person and must be adaptively learned by the machine for each speaker. Simple circuitry is described that learns a word with a single utterance and recognizes it thereafter. The scheme is potentially economical for the spoken equivalent of key entry of data and data inquiry, and a limited vocabulary of commands to the equipment.

Speech is communication. A speaker encodes his thoughts into a sequence of sounds so that the listener can understand what he is saying. Speaking in a given language, he follows certain rules of grammar and pronounces his words in the accepted way. Differences such as rate, pitch, accent, volume, and tone quality have little effect on recognition accuracy because the hearer quickly adapts to the speaker's voice. In addition to the intended message, the listener may receive one or more speaker-related messages, such as the emotional attitude, state of health, age, and sex of the speaker. Even the speaker's place of birth and educational background are often revealed to the attentive listener.

Since speech is man's basic method of communication, it is natural to ask the question: "Why not talk to machines?" We have made some good beginnings. The telephone was considered to be only a rich man's toy when it was invented; now, after a hundred years, it numbers in the hundreds of millions and is found in practically every home and office in the United States and throughout the civilized world. The phonograph, public address system, radio, television, and tape recorder are other examples of mechanical aids to communication between men.

The telephone and radio have broken the distance barrier and the phonograph and tape recorder have allowed sounds to be recorded so that time is spanned and we can hear Caruso sing again. Now man seeks to bridge the man–machine communication gap. Already it is possible to program the computer to respond with prerecorded voice messages, and work is in progress to improve memory utilization by speech synthesis.[1]

Automatic word recognizers must use some form of speech analysis. The pioneering efforts of engineers concerned with bandwidth compression for efficient speech transmission have laid the foundation for work in both speech analysis and speech synthesis. Homer Dudley of the Bell Telephone Laboratories developed one of the first speech analyzers, known as the Vocoder.[2] Early attempts at word recognition, the problems and pitfalls, are portrayed in an excellent survey article by Lindgren.[3] A survey by Hyde[4] covers more recent work.

The voice-operated typewriter has been the goal of inventors in the United States and elsewhere. However, they have not been successful because there is no one-to-one correlation of sound to the printed symbol. Several digit recognizers have been built that operate quite well for the speaker to which they are tuned. Computers have been used as speech analyzers and also as recognizers in conjunction with a hardware analyzer.[5] The problems of connected-word recognition are being attacked by the use of modern digital computing techniques.[6]

It is the purpose of this article to describe an automatic word recognizer using an adaptive memory technique for quickly adjusting to the speaker and the language spoken. A small, but arbitrary, library of words is used. Potential applications of such a device include aid to paraplegics (to control bed, lights, etc.); parcel sorting (registration of Zip codes while handling packages); air traffic control (to switch radar range, etc.); aid to jet pilots or astronauts (to adjust radio receiving channel, etc.); and remote inquiry via telephone or radio link to computer.

Speech has much to offer as an input to a system. It leaves the hands free for other activities, and it can be used in the dark and does not require a writing implement. Some form of feedback—visual, aural, or even tactile—can be used to assure correct entry.

Physical properties of the acoustic signal

Homer Dudley[2] has provided a perceptual model of speech that is useful. He suggests that speech consists of modulated carriers upon which slow variations of resonance and intensity are impressed. The two carriers are the "buzz" of the vocal cords and the "hiss" of the breath escaping through a constricted portion of the vocal tract. The vocal cords vibrate at a relatively fixed rate for conversation, but small changes in the frequency of vibration or "pitch" may convey information. The rising inflection of the voice at the end of a sentence to denote a question is an example. The relatively slow variations in harmonic resonance, pitch, and intensity are the information-bearing elements of speech.

There are two main classes of speech sounds: unvoiced and voiced. A possible third class is formed by completely interrupting the breath outflow so that pressure can build up and then be released explosively. The silence preceding the "sound burst" is an obvious feature of the word, since the sound is continuous except for these breaks within the word. The "stop" consonants are p, t, k, and b, d, g; the first three are unvoiced and the last are voiced. The "silence" before the voiced stop consonants is not complete, since the vibration of the vocal cords is present but subdued. All stop consonants are characterized by a sudden burst of intensity following the pressure release. Both p and b are formed by closing the lips; t and d by the tip of the tongue and hard palate; and k and g by the back of the tongue and soft palate.

Other unvoiced consonants are formed by partially restricting the free flow of air so that a hiss is produced. The soft hiss formed by the space between the upper front teeth and the lower lip is the /f/ sound. The harsher sound formed by the tip of the tongue against the hard palate is the /s/ sound. A lower-pitched hiss results when the tip of the tongue drops and the opening is less restricted to form the /sh/ sound.

A fourth class of sound results from mixing the vibration of the vocal cords with the turbulence of the constricted air passage. Thus an f becomes v and s becomes z by adding voice.

The voiced sounds outnumber the unvoiced sounds by about five to one in the English language, the vowels being the largest single class of sounds. These sounds result from air passing through the vocal cords, which vibrate in response to this action in such a way that a triangular or sawtooth-shaped pressure–time pattern is produced. Such a forcing function is rich in harmonics and serves to excite portions of the vocal tract so that some harmonics are strengthened and others are diminished. Thus, the wavefront issuing from the mouth and nose is a complex waveform varying with each change in position of the organs of speech.

Through years of practice, we have all become experts at making acceptable speech sounds to convey messages to the listener. This involves precise control of the vocal cords, throat muscles, soft palate, tongue, jaw muscles, and lips, and, of course, the diaphragm, chest, and other muscles concerned with breath control. Fortunately, we do not think of all these actions consciously, so a thought becomes a word or phrase with little effort or attention.

The ear is a marvelous mechanism connected to the powerful analyzer that is the brain. Since theories of hearing are quite varied, we shall not attempt to analyze them. We shall say only that there is general agreement that the ear and the brain together perform a frequency analysis of the complex speech waveform. The net result is a demodulation of the impressed intelligence so that the speaker's message is received. Likewise, we leave all questions of grammar, phonetics, semantics, and syntax unanswered, since these will not enhance our understanding of the word recognizer that operates in what is essentially a context-free environment.

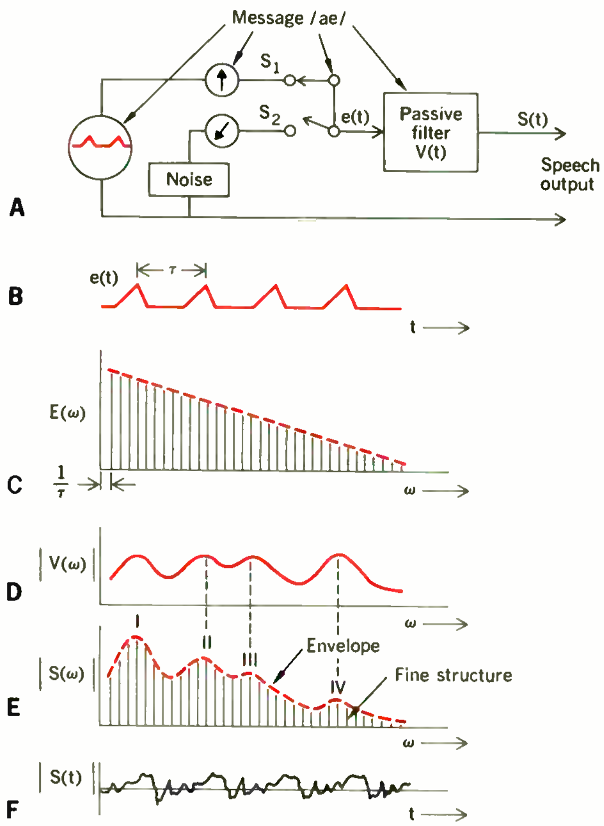
Production of the acoustic signal

It will be helpful to think of the acoustic signal of speech as being the output of the electrical network shown in Fig. 1A. Here the passive filter represents the resonance conditions of the vocal cavities. The excitation function $e(t)$ is either a periodic sawtooth waveshape representing the buzz from the vocal cords or a noise source, the hiss of the unvoiced sounds. Both sources are applied for combination sounds such as /V/ or /Z/. In the example, the sound /ae/ is being produced as part of a message. (This is the sound of the short a in the word *at*.) The excitation function is the sawtooth as shown in Fig. 1B. The period T is about 8.3 ms for a male voice having a fundamental frequency of 120 Hz. For the purposes of the example, it is assumed that the frequency is constant, so we may perform an analysis in the frequency domain.

The line spectrum of the excitation function $E(\omega)$ is shown in Fig. 1C. The spacing is $1/T$ or 120 Hz, so harmonics appear at 240 Hz, 360 Hz, 480 Hz, etc. There is a gradual falling off in intensity as frequency increases up to about 4 kHz, at which point the effect of the harmonics is negligible. Within this frequency range there may be as many as four "poles" or points of resonance in the frequency response of the vocal cavities. Figure 1D illustrates the adjustment for producing the sound /ae/. The output of the passive filter, $S(\omega)$, is shown in Fig. 1E. This is a composite line spectrum having the fine structure of the excitation function and an envelope having the general shape of the passive-filter frequency response with the overall negative slope of the excitation function. Peaks are numbered to correspond to the conventional method for numbering speech features to be described later.

Going from the frequency domain to the time domain,

FIGURE 1. Production of the acoustic signal. A—Block diagram. B—Excitation function. C—Line spectrum. D—Frequency response. E—Composite line spectrum. F—Acoustic signal.



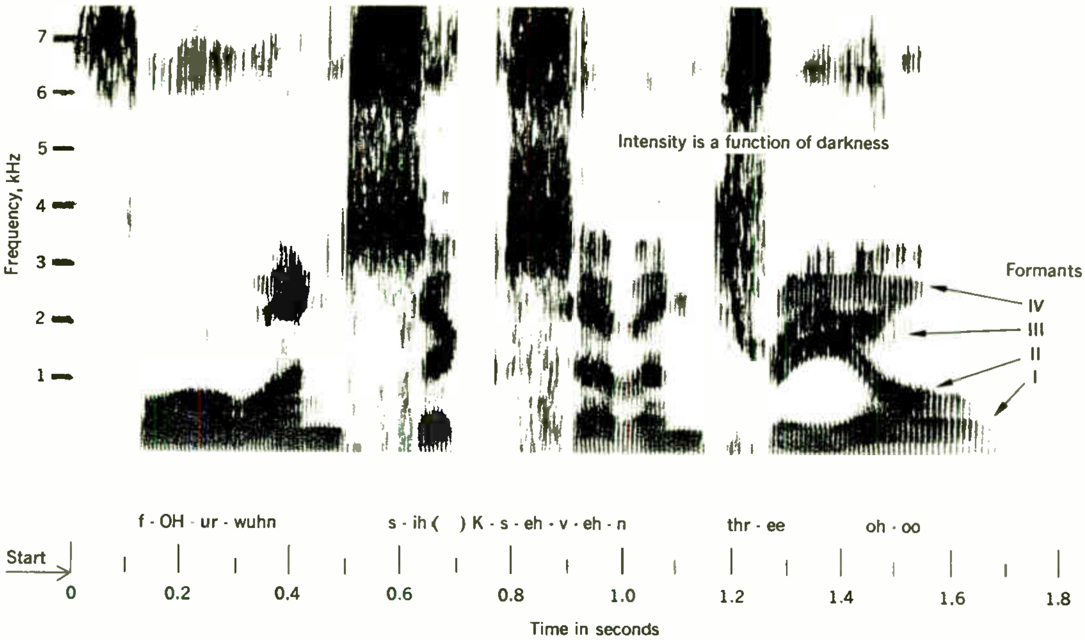


FIGURE 2. Sonogram of male voice.

we see the acoustic signal $S(t)$ as shown in Fig. 1F. This is a representation of the pressure–time function of /ae/ as it would appear on an oscilloscope connected to a microphone actuated by a male speaker. The excitation function is visible for this "open" or flat-sounding vowel. The harmonic fine structure and the four poles of the frequency response contribute to the peculiar structure of this complex waveform in the time domain. Of course, for this example we have selected a short period in time when the vocal cavities are relatively fixed and the pitch and intensity relatively uniform. In practice, the acoustic signal varies rapidly from the random fluctuations of the noiselike unvoiced consonants through the many complex shapes of the vowel sounds as they are produced by the movements of the vocal mechanism in speech. Thus, the acoustic signal falling on the ear or microphone is extremely complex and variable in the time domain. Some form of spectrographic analysis is required to bring out identifiable speech features.

Spectrographic analysis

An important tool in speech research has been the sound spectrograph,[7] developed at Bell Telephone Laboratories in 1945. It provided visible evidence of the resonances of the vocal tract that produce patterns of energy concentration in the frequency domain known as "formants," which have since been used in speech analysis and synthesis. The successful use of formants to produce intelligible speech has demonstrated that they are important information-bearing elements.

Figure 2 is a reproduction of a sonogram produced on the sonograph (a spectrum analyzer manufactured by Kay Electronics Co.). A sonogram is a three-dimensional display of energy (darkness of line), frequency (ordinate), and time (abscissa). In this particular sonogram a relatively wide bandwidth of 300 Hz was used to eliminate the frequency-fine structure of the harmonics in the voiced portions and to bring out the time-periodic nature of the voice fundamental as evidenced by the vertical striations. These show that the pitch varied from about 130 Hz at the beginning to about 110 Hz at the end of the phrase, where the voice pitch dropped. Thus, the glottal impulses or excitation function for the voiced portions are clearly revealed. A narrow bandwidth around 50 Hz would have made the harmonic line spectrum visible, but poor time resolution would have eliminated the glottal impulses. This tradeoff in time and frequency resolution is similar to the Heisenberg "uncertainty principle" in that the product of bandwidth and time resolution limit is a constant.

By eliminating the harmonic-fine structure and improving the time resolution, the true nature of the formant action is revealed. The relatively horizontal, broad bands of energy below 4 kHz show the action of the peaks of the envelope function as the frequency response of the vocal cavities varies during the utterance of the phrase *416730*. The formants are numbered as shown at right using Roman numerals (compare Fig. 1D).

Many speech features are visible in the sonogram of Fig. 2. The phrase starts with the sound of /f/ in *four*, which appears as high-frequency energy near 7 kHz and above. The amorphous structure contrasts with the vertical striations of the voiced /oh/ that follows. The close conjunction of formants I and II at the beginning of /oh/ widens slightly and then closes near the 0.3-second mark to indicate the end of *four* and the beginning of *one* (wuhn). *One* is characterized by the upward glide in formant II and the addition of formants III and IV between 0.3 and 0.4 second followed by the low-frequency sound of /n/, which lasts until 0.5 second. At this point voicing ceases and the strong sibilant character of the /s/ in *six* is shown by the heavy concentrations of "noise" energy around 4 and 6 kHz and higher. This is followed by the short vowel sound /ih/, a period of silence around 0.7 second, the narrow vertical line for the stop consonant /k/ at 0.75 second, and the final /s/ sound of the x in *six*.

Between 0.8 and 0.9 second there is an example of elision, since the end of *six* is also the beginning of *seven*. Then the vowel sound /eh/ appears on either side of a gap in the formant structure representing the subdued /v/ sound. The /n/ at the end of *seven* is similar to the /n/ at the end of *one*.

The /th/ sound at the beginning of *three* differs from the previous /s/ sounds in the distribution of energy. Also, at around 1.2 seconds note the formant-like glide of the rolled /r/ sound, which alternates in intensity as it merges into the /ee/ sound. The distinctive nature of the vowel /ee/ sound is shown by the wide separation of formants I and II.

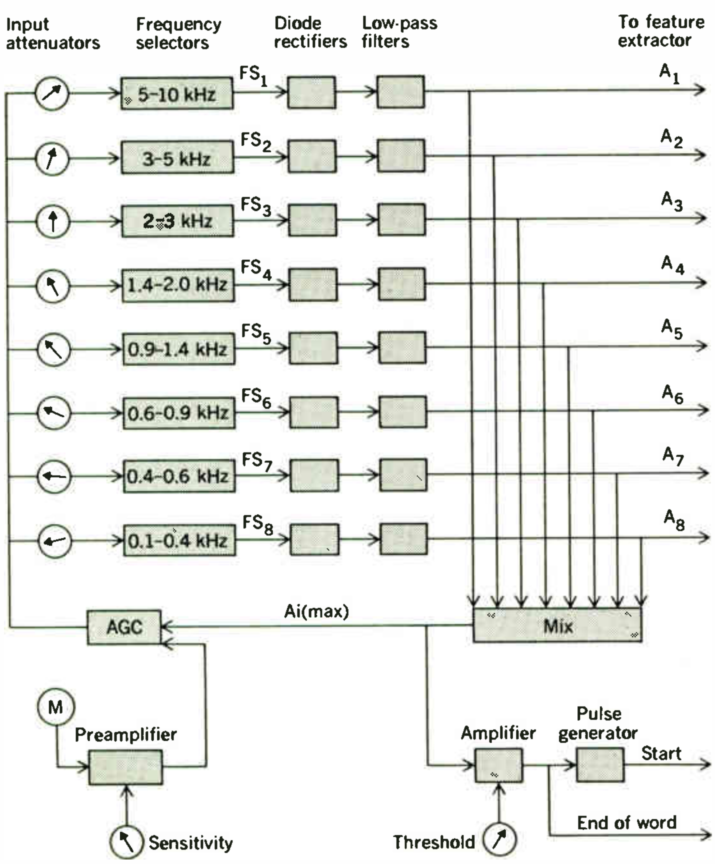
The transition from *three* to *oh* is marked by the downward bending of formant II toward formant I, which rises slightly to produce the /oh/ sound. Formant II fades out before formant I and the phrase ends on a faint /oo/ sound, which is rapidly dropping in the fundamental frequency or vocal pitch as shown by the widening intervals between successive glottal impulses.

From this single sonogram a number of speech features have been noted: fricative /f/—broadband noise energy at 7 kHz and above; sibilant /s/—noise energy at 3 kHz and above with a "window" at 5 kHz; silence—before stop consonant, no visible energy; voice pitch—glottal pulses per time interval, 3 kHz and below; steady formants—/N/ in *one* and *seven;* formant upward glide—*one;* downward glide—*three-oh;* close formants—/oh/ in *four;* wide formants—/ee/ in *three*.

Actually, there is such a wealth of information that the eye must select certain dominant features from the many confusing details. Nevertheless, the features are much more accessible than they are in the complex waveform of the acoustic signal.

In recent years, the general-purpose computer has been programmed to provide spectrographic information directly from the acoustic signal. The flexibility of the computer implementation is an advantage in speech research. However, like the sound spectrogram, it provides more detailed information than we have found necessary to recognize a limited vocabulary of individual words.

FIGURE 3. Speech analyzer.



Individual word recognition from a single speaker

The large amount of information contained in a complete time-based spectrographic analysis of normal connected speech received from random speakers far exceeds the minimum required to recognize a limited number of individual words received from a single speaker. The work reported here is the reduction to practice of a recognizer with that limited capability and it is able to function effectively using only the signal levels in eight bandpass channels, each of which uniquely covers a portion of the audio-frequency range. The rectified and low-pass filtered channel outputs are continuously compared with each other. An analog dynamic threshold scheme selects up to four of the outputs as significant and produces binary one levels for the significant channels. A nonlinear time slot sequence is initiated at the start of a word, arranged so that the duration of the slots increases as the sequence proceeds. Thus each word can be represented by a characteristic two-dimensional pattern of ones and zeros formed by a matrix of the time slots and the frequency channels. If at any time during any time slot a particular channel is rated as significant, a one is placed in the appropriate matrix position.

The machine has a training mode in which the pattern of a spoken word is placed in a special memory plane or template. There are as many templates as the memory capacity of the machine. After each template is thus trained to its word, the machine can be put in its recognition mode. In the recognition mode the pattern acquired from a spoken word is compared in parallel with the stored pattern in all templates and the best match selected.

The features believed to be novel in our approach are the nonlinear time base,[9] the analog dynamic threshold, significant channel selector,[10] and the template circuitry.[11]

Speech analyzer

A practical speech analyzer for an automatic word recognizer is shown in Fig. 3. The acoustic signal is converted by the microphone *M* to electric signals that are amplified in the preamplifier and controlled by automatic gain control (AGC) to provide signals with relatively uniform overall amplitude to the frequency selectors. Input attenuators are adjusted to set the overall gain of each selector so as to offset the negative slope of the envelope function that was noted in Fig. 1E. The sensitivity control is adjusted to eliminate background noise.

The frequency selector ranges are designed to give optimum coverage of the frequency spectrum from 0.1 Hz to 10 kHz. First, a broadband frequency selector covers the range from 5 to 10 kHz containing the high-frequency noise energy of the fricative and sibilant sounds. This selector (FS_1) uses a low-pass filter and a differential amplifier to secure a broad high-pass filtering action with a sharp cutoff at the 5-kHz window. The next selector (FS_2) is a moderately broad bandpass filter of standard design covering the 3–5-kHz frequency range. From the sonogram of Fig. 2, we saw that this was the region with the concentration of noise energy for the sibilant sounds. Selectors FS_2–FS_7 have ranges that are equally spaced when plotted on a scale representing the logarithm of the frequency. Thus the ranges are packed more closely in the lower half of the formant region with most of the formant action, five of the eight selectors covering the frequency spectrum from 0.4 to 3 kHz. The lowest frequency range, 0.1–0.4 kHz, is covered by FS_8, which has a broad bandpass to encompass both male and female voice fundamental pitch frequencies.

Having quantized the frequency spectrum into bands broad enough to remove the harmonic fine-line structure, the selector outputs FS_1–FS_8 are rectified and smoothed to detect the envelope function. Thus, a short-time integration is achieved. The outputs from the low-pass filters are slowly varying dc levels whose amplitudes at any given time correspond to the envelope function. Since the input attenuator adjustments compensated for the negative slope of the composite line spectrum in Fig. 1E, the shape of the envelope defined by amplitude functions A_1–A_8 approximates the curve in Fig. 1D. This curve is the frequency response of the passive filter that represents the vocal-cavity resonances. Therefore it is evident that the speech-analyzer outputs, A_1 to A_8, produce frequency-quantized envelope amplitude functions describing the changes in a speaker's vocal-cavity resonances in real time.

The speech analyzer outputs are mixed together in a diode positive OR circuit to provide the control voltage for the AGC. During an utterance, the output having the highest amplitude controls the level of the input to the frequency selectors so that none are overdriven. At the beginning of a word, any analyzer output that rises above a threshold generates a start signal for the time-base generator. The amplitude function mix also controls the end-of-word detection within the time-base generator.

Feature extractor

The principal function of the analyzer outputs is accomplished through the feature extractor shown in Fig. 4. This device performs the function of the eye as it scans a sonogram looking for features. Just as the eye takes note of differences in the darkness of various parts of the sonogram of Fig. 2, so the feature extractor compares the analyzer outputs A_1–A_8 against threshold voltages E_1–E_8 that are derived from a resistor network. Each threshold voltage tends to follow its own input, being held to a voltage no lower than a few tenths of a volt below the input voltage. Through the resistor network it affects all the other thresholds, with the greatest effect upon the immediate neighbors. Thus, the local maxima in the envelope function are effective to produce outputs from the amplitude-comparison circuits and at the same time prevent outputs from neighboring units having inputs of lesser amplitude. The effect of the resistor network is to produce a "floating" or self-adjusting threshold that brings out the poles of the envelope function regardless of the absolute amplitude. A constant-current source limits the maximum number of on units to four. The threshold voltage, E_0, provides an absolute minimum threshold that must be exceeded before any unit can operate. The outputs of the amplitude-comparison units are applied with units having hysteresis and shaping so that the final outputs, H_1 to H_8, are well-shaped, jitter-free, square-wave pulses of standard amplitude. These are the inputs to a matrix of storage units that store the envelope information with reference to a time base.

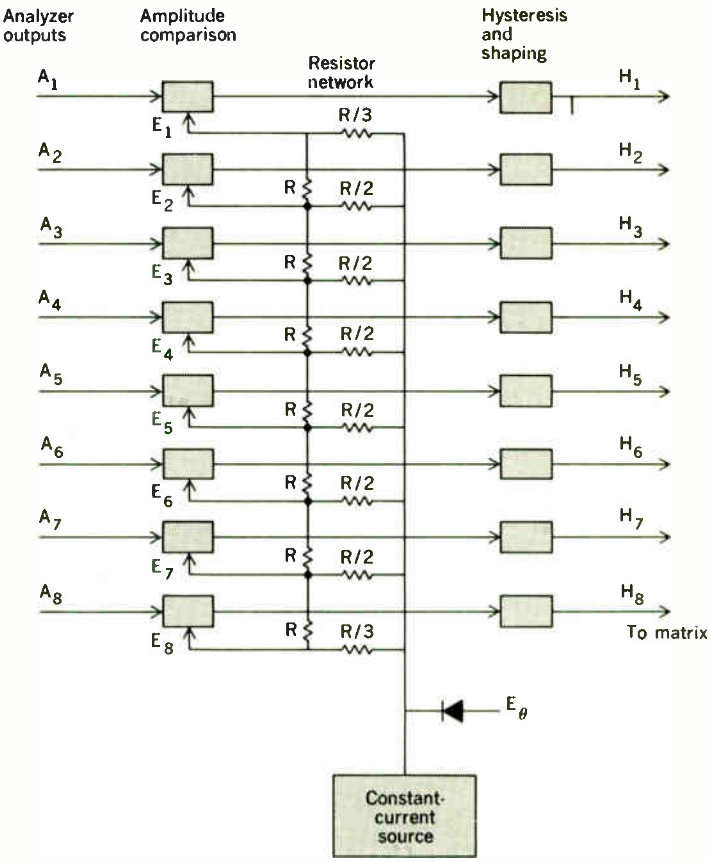
Nonlinear time base

Speech features are stored with reference to a nonlinear time base. Nonlinearity is used to achieve a desirable compression of information while removing the effects of uncertainty in time position. In situations where discrete words are used, it has been observed that sounds close to the start of the word are more consistent in timing than those nearer the end. When sampling is done at regular intervals, the variation in position seems to increase linearly with distance from the beginning of the word. At a certain sampling rate, the variation or "spread" may be as follows:

Sample no.	1-2-3-4-5-6-7-8-9-10-11-12-13-14-15-16-17-18-19-20-21
Spread	-,---,----,-------,------------,---------------
Time units	1 2 3 4 5 6

By sampling at a rate that starts at the given sampling rate but constantly slows down with time, the number of time units in each succeeding time slot can be made to

FIGURE 4. Feature extractor.



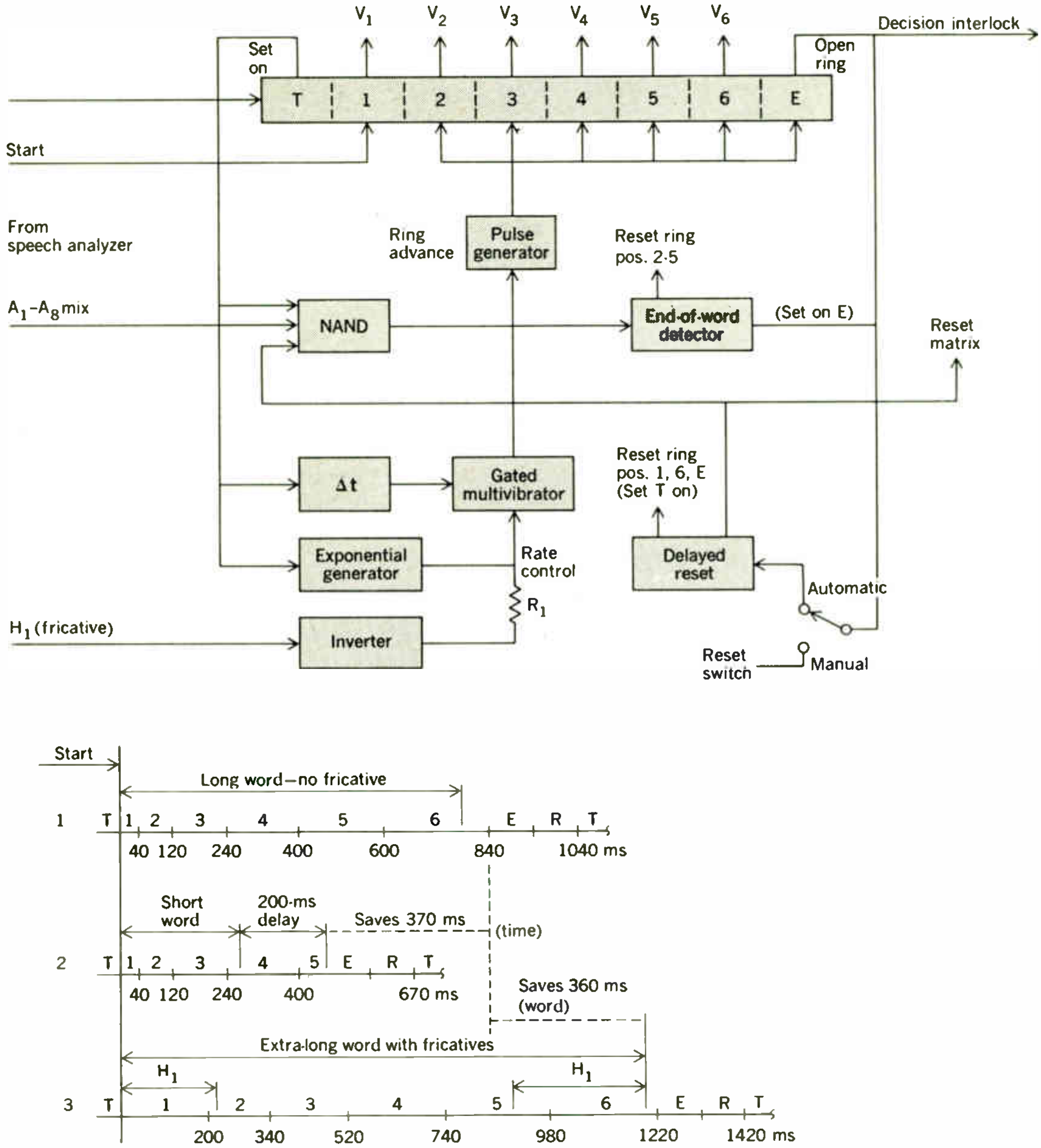


FIGURE 5. Nonlinear time base.

increase linearly:

Time slot no.	1	2	3	4	5	6
Time units	1	2	3	4	5	6
Elapsed time units	1	3	6	10	15	21

Thus, each successive time slot widens to receive the expected variation of the central feature. Of course, features still may appear in two time slots whenever they occur near a slot boundary but this is preferable to having them spread over five or six slots or sampling positions. Also, there is a tendency to cluster the final features of a word, but this is offset by the speaker's natural tendency to draw out or prolong the endings of words and to be crisp and precise with the beginning sounds. The net effect is a time compression and normalization of speech features with some blurring of detail that is not serious. In fact, there is reason to believe that some smearing of a quantized pattern is beneficial. This can be tested by looking at a picture that has been quantized into squares of uniform gray tones. At first, only a mosaic of various shades of gray is seen, but when blurred by distance or looked at through the eyelashes, the picture suddenly appears.[12]

The nonlinear-time-base schematic is shown in Fig. 5. For discrete words of one or two syllables an open ring of eight units has proved satisfactory. The ring is set on *T* (talk) and the start signal from the speech analyzer turns on unit 1, which turns off *T* and gates units 2. The ring then advances through 2, 3, 4, etc., until *E* (end) is reached. As each ring position 1 through 6 is energized in turn, a corresponding line, V_1–V_6, is raised to gate a matrix column to store the feature or features present during that particular time interval.

The start signal starts the exponential generator by turning off the *T* position and, after a short delay, it starts the gated multivibrator. The rate control voltage drops

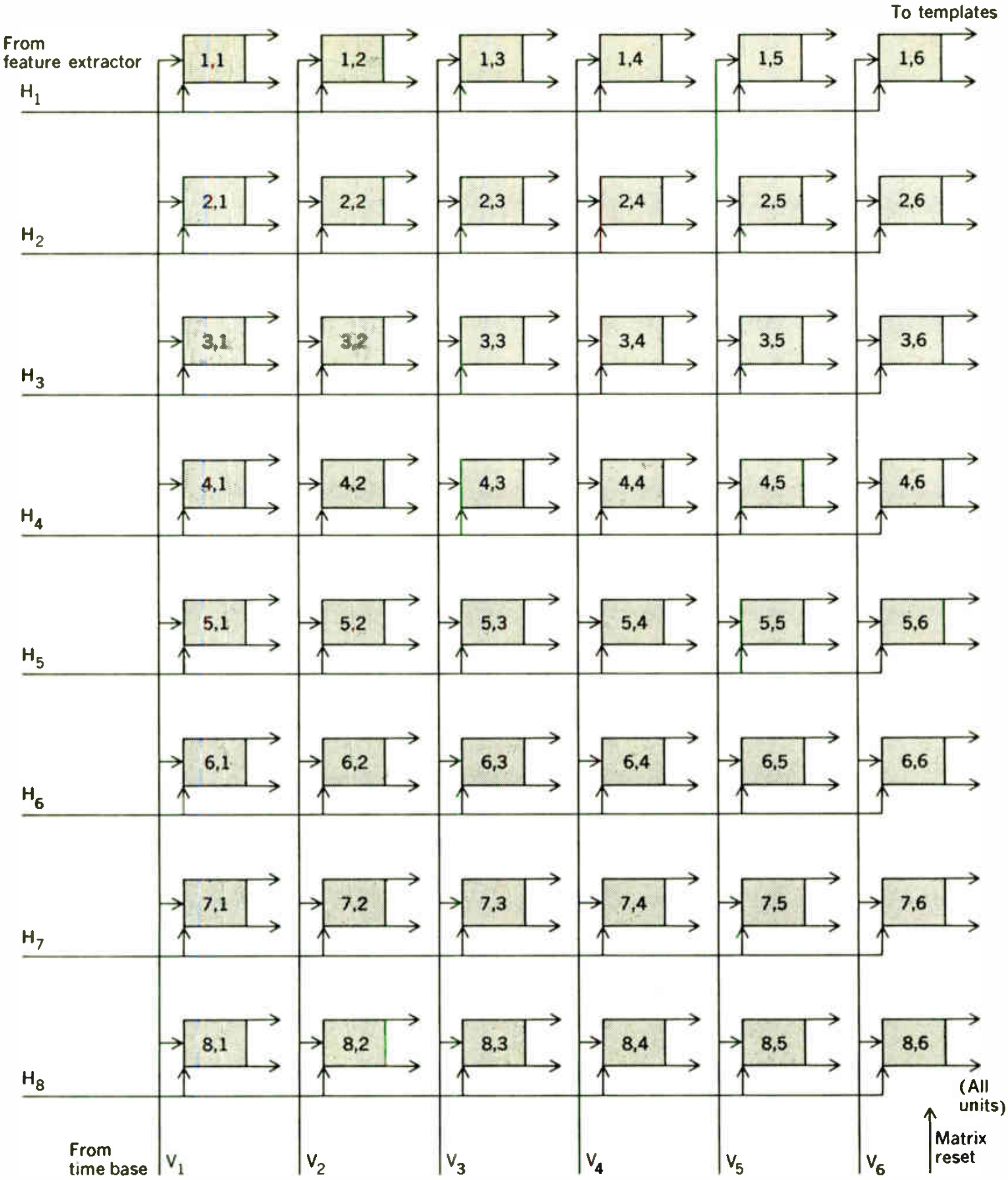


FIGURE 6. Storage matrix.

steadily for a long word with no fricative sounds, so each succeeding time interval lengthens by the chosen unit of time, 40 ms, in the example. The delay adds to the gated multivibrator turn-on time to make the first 40-ms interval. Succeeding pulses appear on the ring advance line at 120, 240, 400, 600, and 840 ms. At this time, E turns on and starts an automatic reset cycle. A built-in delay of 100 ms allows E to remain on for this time duration to assure enough time for the decision interlock. Then the reset interval R begins and all ring positions and matrix storage positions are reset. After another 100-ms interval, T is set on and the cycle can begin again. This is the timing for the long word with no fricative, as shown in the first example.

In the second example a short word of about 270 ms is assumed. Since it would be wasteful of time to go through the full ring cycle for this short word, an end-of-word detector is energized after a short delay. Normally, the end-of-word detector is held off by one of the three inputs to a NAND (negative AND) circuit as shown: T, A_1 to A_8 mix, matrix reset. After the matrix is reset and T is turned off, only the A_1–A_8 mix controls the end-of-word detector. A delay of 200 ms is built in to prevent operation on the short silences before stop consonants. However, if the time that A_1–A_8 is off exceeds 200 ms, it is assumed that the word has ended and E is set on and intermediate ring positions are reset off. In the second example, for the 270-ms word, the 200-ms delay brings the time from T to E to 470 ms rather than the 840 ms for the full cycle. Thus, 370 ms of time is saved.

In the third example a word begins and ends with fricative sounds. Inasmuch as these sounds are quite

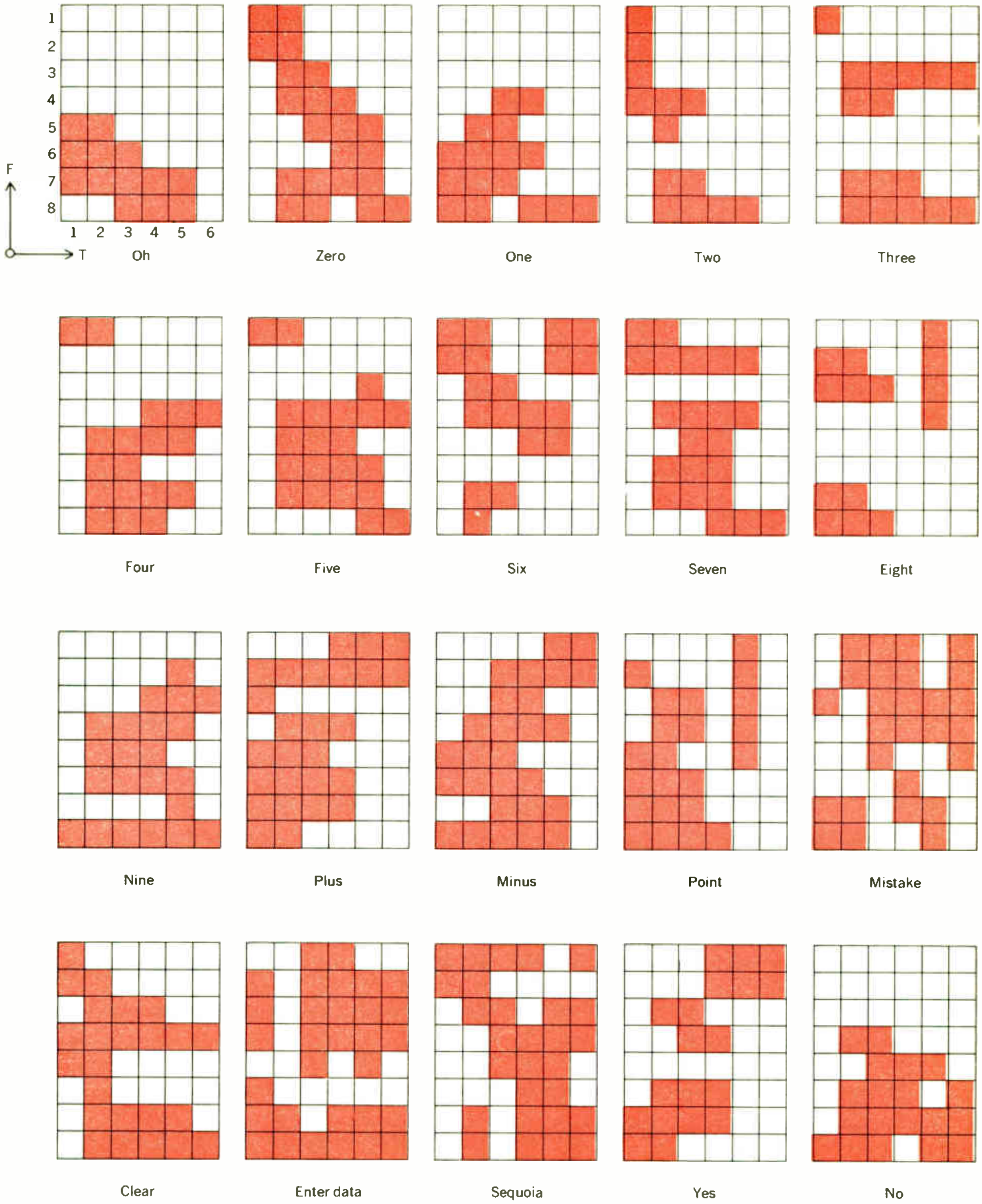


FIGURE 7. Word patterns.

long and the variability in duration is high even for beginning sounds, it is necessary to incorporate an additional control to slow the multivibrator rate when a fricative sound is present. The H_1 input to an inverter lowers the rate control to a voltage determined only by R_1 and not by the exponential generator. At the end of the fricative sound, the voltage returns to the level set by the exponential generator. Thus, the first interval is lengthened to about 200 ms and succeeding time intervals are controlled by the exponential voltage generator. The final H_1 or fricative input does not change the rate of the multivibrator since it has already slowed to a point beyond the rate set by the fricative interlock. In this example a word about 1.2 seconds in total length is accommodated by the feature-storage time base, saving about 360 ms of word duration. Also, the pattern is more

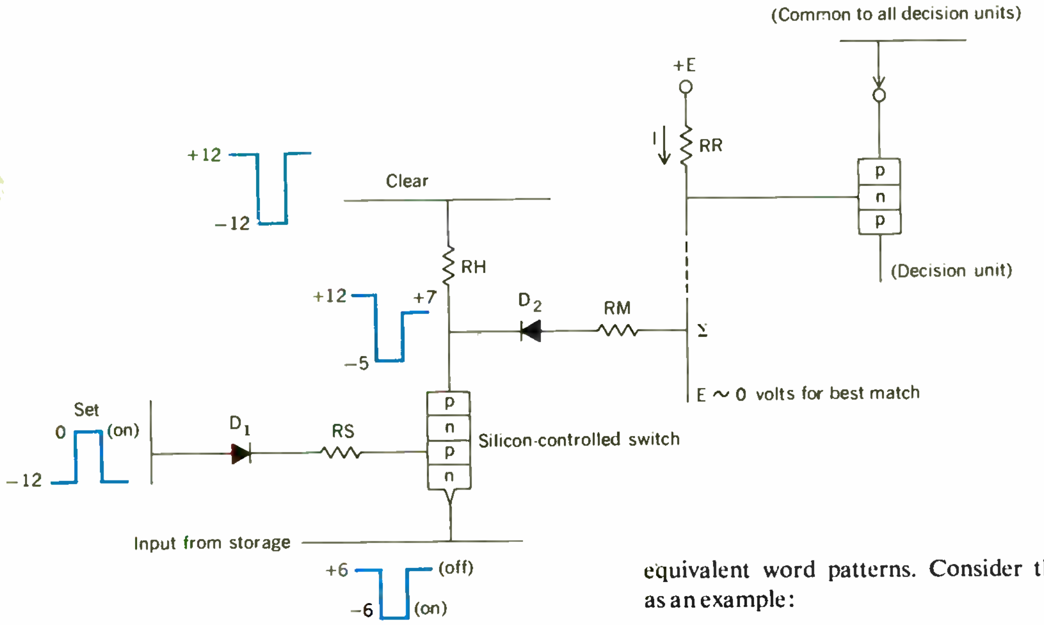


FIGURE 8. Electronic template unit.

compactly stored at the beginning of the word because the fricative occupies only one or two time slots rather than three or four. The resultant compression permits efficient use of available storage-matrix positions.

Storage matrix

Eight output lines from the feature extractor and six lines from the nonlinear time base are the horizontal and vertical lines, respectively, for the storage matrix shown in Fig. 6. A storage unit is located at each intersection and is identified by the row and column numbers of the lines to which it is connected. A coincidence of input signals on these lines sets the storage unit on. For example, an input on H_1 during time slot V_1 results in turning on storage unit 1, 1. Thus, a word pattern of frequency versus time is formed. "On" units present a pattern of word features that bears some resemblance to the spectrographic analysis of Fig. 2. "Off" units are significant too, so that both direct and complementary outputs are used. In our example, there are 48 storage units and 96 outputs to the adaptive electronic templates. The storage units are symmetrical flip-flops with indications so that word patterns are visible. Once set on, a pattern is stored until reset. During adaption of the templates the reset control is manual, so patterns may be used only after inspection. For automatic word recognition, the word pattern is automatically reset after a decision has been made. Typical word patterns are shown in Fig. 7.

Word patterns

The 20 word patterns shown in Fig. 7 are similar to those made by the spectrographic analysis previously described. It must be remembered, however, that both frequency and time scales were linear in the sonogram, but in Fig. 7 these are both nonlinear. Also, these words were spoken separately but the words in the sonogram were connected. Moreover, pitch is not evident because of the coarse time quantization. Nevertheless, similarities may be seen between the words of the sonogram and the equivalent word patterns. Consider the pattern for *six* as an example:

Column 1. H_1 and H_2 denote the sibilant /S/ sound.

Column 2. H_3, H_4, H_7, H_8 represent the beginning of /ih/ mixed with H_1 and H_2, which persisted into the beginning of the column 2 time slot.

Column 3. H_3, H_4, and H_7 show a change in the /ih/ sound (notice that the stop consonant silence is lost).

Column 4. H_4 and H_5 represent the /K/ sound that must have occurred near the boundary between time slots 4 and 5 because it appears in both columns.

Column 5. H_1 and H_2 reappear with H_4 and H_5, showing the end of the /K/ sound and the beginning of the /s/ sound.

Column 6. H_1 and H_2 alone represent the final sibilant /s/ sound of *six*.

Thus, all the useful features seen in the sonogram representation of *six* are present but the silence before the /K/ sound. This particular feature appears in the word pattern for *eight* before the final /t/ sound. In general, salient features such as the upward glide in *one;* the downward glide in *oh;* the wide separation of formants in *three;* the fricatives in *four* and *five;* the /n/ sounds of *one, seven, nine, point, no;* and other formant locations and transitions such as the /ah-ee/ sound of i in *five, nine,* or *minus* are recognized quite easily by the practiced eye. The broadband frequency analysis, the nonlinear time base, and the amplitude normalization of the feature extractor combine to produce separable word patterns that are consistent for an individual speaker.

Adaptive electronic template

Different dialects and speaking habits as well as built-in personality differences cause variations in the patterns produced by one speaker as compared with another. Therefore, some form of adaption to the speaker is indicated. If the time needed for adaption is short, it will not be inconvenient for the speaker occasionally to update the word-recognition device to provide a good reference base for recognition. It is also a way to change a word in the library or to add to or subtract from the word list. Of course, adaption allows an entirely new library of words to be entered—a different language, for example.

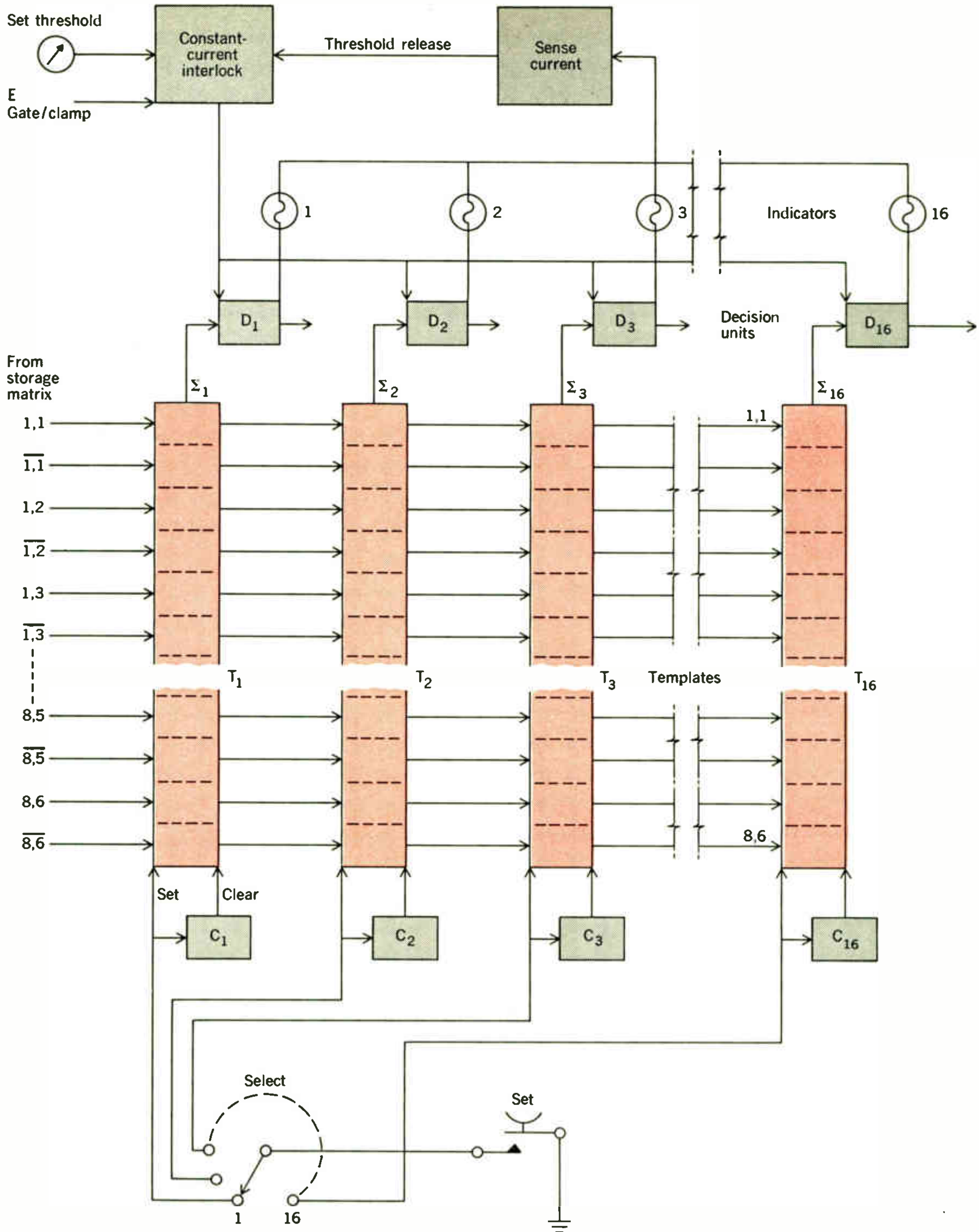


FIGURE 9. Adaptive electronic templates.

The adaptive electronic template system to be described has extremely short "training" times compared with systems using error-correction routines, because the iterative procedures of these systems are not required. In general, only one pass through the word library is needed for the adaptive electronic template system. A check pass with a second entry of any troublesome word usually gives best results. The whole operation takes two or three minutes for a library of 16 to 20 words as compared with several hours, or even days or weeks, in the former systems. Thus, the speaker is better able to update the library or correct a malfunctioning word. Furthermore, there is no need for permanent, nonvolatile storage in the electronic template system. Instead, simple, low-cost, highly reliable binary switches are used.

The operation of the electronic template can be understood by referring to the unit schematic shown in Fig. 8. During adaption, the coincidence of a positive set pulse and a negative (on) input from a feature storage unit turns the silicon controlled switch on. The resistance of

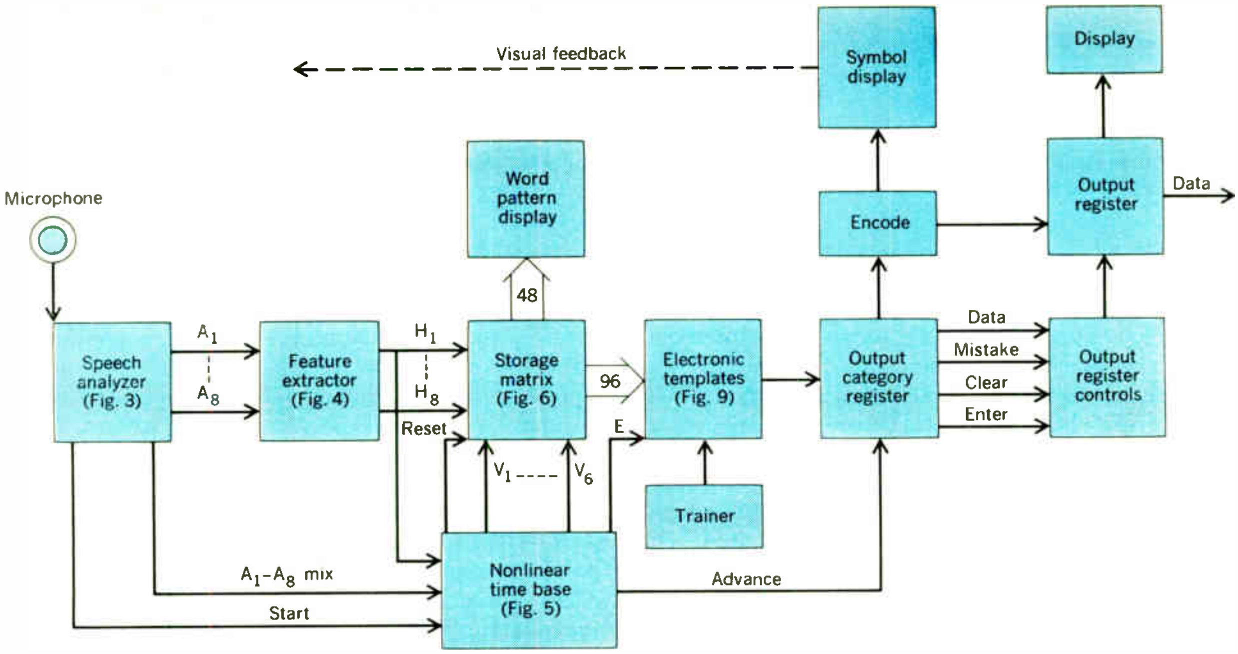


FIGURE 10. Automatic word recognizer.

the switch drops to a low value and the conductive condition is maintained by holding current flowing from the clear line through RH, when the input line is at the (off) upper +6-volt level. Once it is turned on, the switch anode follows the input so that current flows in measuring resistor RM when the input is at the −6-volt level.

For matching operations, units that are conductive, but have inputs that are off, do not affect the voltage at the summation line because D_2 diode is back-biased. Also, units that are not turned on by the adaptive process do not contribute to the summation because the silicon-controlled switch does not conduct. Thus, a unit of current is added to the total current affecting the summation line only when there is a coincidence of a previously turned on switch and a present "on" input from feature storage.

An electronic template consists of a group of 96 basic adaptive units operating on a common summation line connected to a decision unit. That template having the nearest match between the pattern of inputs from feature storage and the pattern of conducting silicon controlled switches will have maximum current flowing in the summation line. This causes the maximum voltage drop in the reference resistor RR and thus the template output voltage is most negative for the nearest match. All other template outputs will be above this voltage and will represent no match.

The organization of the adaptive electronic templates is shown in Fig. 9. Inputs from the storage matrix are connected to all 16 templates in parallel. Each template consists of 96 of the units described in Fig. 8 because both direct and complementary outputs of the 48 storage matrix units are used. Thus, each word pattern is a certain combination of "off" units and "on" units in the storage matrix, so some complementary outputs and some direct outputs are active; however, the total number of active lines into the templates is always half of 96, or 48.

During the "training process" a word is spoken and a pattern is stored in the storage matrix, so when template 1 is selected by means of the *select* switch and the *set* switch is closed, those individual template units having active input lines from the storage matrix will be set on in T_1. Similarly, a second pattern may be set into T_2, etc. Individual clear units for each template automatically turn off all individual template units for the template elected just before the new entry is made. With this system it is impossible to superimpose patterns. Furthermore, substitutions and corrections may be made without clearing all templates.

It is obvious that immediately after adaption the input pattern exactly matches the pattern stored in the template, and therefore maximum current flows in the summation line for the template just adapted. Thus, the output from the corresponding decision unit may be used to check the validity of the adaptive procedure. When all word patterns are stored in the desired template locations, the adaption is complete. The words may now be repeated in any desired order, and the accuracy of recognition is noted by observing the decision unit outputs. Any word that is not recognized because of a faulty match may be easily corrected and rechecked until consistent recognition is obtained. The adaption and checking may be done with a relatively high decision threshold to assure optimum performance during later recognition.

The constant-current interlock provides sufficient current to operate only one of the decision units D_1–D_{16}. As a constant-current source, it also has the property of a variable impedance, so the common line to the decision units closely follows the most negative voltage at the decision unit inputs. In so doing, it cuts off all other inputs, unless two are exactly equal in voltage level. In this case, the current provided by the constant-current interlock is insufficient to operate either unit and the equal condition causes both patterns to be rejected. This is the preferred operating condition, since dual outputs would be confusing. However, the constant-current generator may be adjusted to allow two or more outputs, if this serves any purpose. In any event, the decision is subject to the adjustment of the threshold and the gate from the time-base ring E position.

The gate prevents any decision from being made until the time-base ring has reached E, which denotes the end of the word—or at least that the entry into the storage matrix is complete. When not in the E position, the gate clamps the interlock line to a voltage so negative that no summation line can reach this level. Thus, no decision is permitted until the time-base ring is at E, then the interlock is allowed to rise to a level governed by the threshold adjustment. When this adjustment is most negative, a good match between the incoming pattern and a stored pattern in a template will cause a decision unit to operate, but a poor match will be rejected; that is, no decision will be made for a poor pattern, even though it is the nearest match. The threshold may be adjusted to accept patterns having any desired "closeness of fit." However, this control would create a decision problem, were it not for the threshold-release circuit.

The threshold release is designed so that a full output is secured when the threshold is just barely reached. Since the threshold is continuously adjustable, this condition is always possible. The alternative would be to quantize the threshold adjustment so that the threshold steps fall between decision levels, since these are also quantized. However, this would place severe limitations on component tolerances and power-supply regulation. The threshold release is a practical solution to the problem. The current for the decision unit indicators is supplied through a low-value series resistor, so an incipient output from any decision unit is sensed and amplified to produce a pulse that raises the voltage on the interlock line. This effectively provides positive feedback to the decision unit and a full output is assured. There is a built-in current threshold for the current-sensing circuit, and thus a certain minimum action must be initiated before the threshold-release action occurs. The threshold is restored by the clamping action of the gate line going off when E is reset because the interlock line is lowered beyond any decision unit input and the decision unit that is on is turned off.

Automatic word recognizer with verification

Now that we have looked at the main constituents of an automatic word recognizer, let us see how these are assembled into a system with a few other standard circuit elements to provide the recognition function. Inasmuch as any recognition system cannot be perfect, means for verifying the recognition process must be used to assure correct output at all times. Audio feedback might be used in some cases but, in general, visual feedback will provide the mast satisfactory means for verification of the spoken word. An automatic word recognizer that utilizes visual feedback is shown in Fig. 10.

Initially, the operator uses manual controls to select and adapt electronic templates to the words in the library. A suitable set of words for data entry to a computer are the digits: *one*, *two*, · · ·, *nine*, *zero*, and the words: *plus*, *minus*, *point*, *mistake*, *clear*, *enter*. This is the minimum set that will allow data entry with hands-off operation. The 16 words may be entered and checked in a few minutes. Once the system has been "trained," operation is by voice control alone.

Words are spoken into the microphone and are analyzed to produce amplitude functions A_1–A_8 describing the envelope of the acoustic signal in the frequency domain. The feature extractor operates on the amplitude functions to produce standard amplitude pulses that represent poles and zeros in the envelope of the speech input signal. At the start of the word, the nonlinear time base begins to operate and the output of the frequency extractor H_1–H_8 is distributed in the storage matrix by timing pulses V_1–V_6. Thus, a word pattern is stored and displayed in frequency and time, as shown in Fig. 7.

At the end of the timing cycle, the E line removes the clamp on the electronic template decision units and the word pattern in the storage matrix is compared simultaneously with all previously stored word patterns in the templates. The word "category" or code number that was assigned during the previous training as a function of the template designation is stored in the output category register for the best match. For example, if *plus* was put into template 11, category 11 is selected whenever *plus* is spoken. This category is encoded into binary 01011 for entry into the output register and the symbol "+" is displayed for operator verification.

Each valid data word is a validation of the previous word. Thus, when the operator sees that the symbol "+" is correct, he says the next word and the recognition of this word as a data word causes the output register control to enter the validated code into the output register. In case of an error, the operator says *mistake*, which is recognized to clear the previous word category and prevent its entrance into the output register. Then the correct word is repeated and verified. After all words of the message have been verified and entered into the output register, the word *enter* is spoken and the message is released from the output register to the data channel. The word *clear* may be used at any time to clear all registers and prepare the system for a new message entry.

Because the automatic word recognizer is a stand-alone device, the computer receiving the data has minimum interruption. Many of these devices can operate as inputs to a common computer collection point since the time to send data from the output register to the computer is a small fraction of the word utterance time. For remote transmission, normal multiplexing and modulation–demodulation techniques can be used.

Preliminary performance data

An experimental prototype of the automatic word recognizer was used to gather preliminary performance data for 13 individual speakers, 11 male and two female. The speakers were laboratory employees with varying degrees of familiarity with the experimental device ranging from none to thorough. Six repetitions each of the 16 words in the vocabulary constituted one test run. The number of runs per speaker ranged from one to 11; one subject performed 11 runs, one performed five, two performed three each, four performed two, and five performed one. A total of 3360 experimental utterances were gathered. The overall accuracy of recognition was 94.2 percent. The substitution error (misrecognition of words) was 3.8 percent and the reject error (omission) was 2 percent.

These results are rigorous in that no retraining was permitted for any experimental run once the 16 words had been entered into the system by a given speaker. In a practical situation, a word giving trouble would be reentered to increase the recognition accuracy.

Analysis of the data showed that performance improved with use and that accuracy of recognition was positively correlated with the user's previous experience. Overall accuracy was 96.4 percent for speakers with some previous experience but 87.9 percent for the first run of inexperienced users. Also, a disproportionate amount of the total substitution error was contributed by the latter, who accounted for 59 percent of the total substitution error with approximately 25 percent of the total runs. For those users who had some experience using the system the overall error was 3.6 percent; 2.1 percent were substitution errors and 1.5 percent were reject errors. Users with considerable experience had recognition accuracy scores of 98 to 99 percent.

Follow-up tests indicated that the average naive subject could be brought up to at least a 95 percent score with no more than a half hour of experience. In some cases, a microphone on a headpiece boom was of considerable assistance but in general the user only needed to become familiar with the device. "Mike fright" has been noted with dictaphones, tape recorders, and public address systems so it is not surprising that there should be a similar effect with the automatic word recognizer.

Problems of connected-word recognition

New solutions are required for the specific problems associated with connected-word recognition. Discrete-word recognizers cannot be easily converted to recognize connected words. The discrete-word model just described requires that the time base be reset after each word. Although this is automatic, the necessary time interval is lacking in continuous speech, because the word boundaries are uncertain and, where there is elision, the next word has begun before the last one is completed. Recognition of connected words must be continuous.

The interaction of speech sounds upon neighboring sounds caused by inertial effects in speech production is a serious problem of connected-word recognition. A given "word" will produce different acoustic signals, depending on the context. In a small library of words such as the one we have considered, the proliferation of word types may not be too serious, but it would be unmanageable for larger libraries. Thus, it seems that a connected-word recognizer will not recognize words as such, but rather some smaller more basic unit such as the syllable or phoneme. We have already referred to one such exploratory program.[6] However, the recognition of smaller units requires a subsequent concatenation of these subunits into words, which is a job for a powerful computer. This is reasonable, since we use a significant portion of that best of all computers—the human brain—to accomplish the task of connected-word recognition.

Conclusions

1. Useful applications exist for an automatic word recognizer with a small, but arbitrary, library of words.

2. An adaptive technique is used to quickly adjust the recognition criterions to the speaker's method of speech production.

3. Words are spoken separately and each is verified before the next one is uttered.

4. Messages are assembled and transmitted after verification to assure accurate output with less than perfect recognition.

5. Results obtained to date indicate that an average of one word out of 20 will require correction for operators having a minimum of a half hour of experience.

6. Powerful computing techniques are needed to recognize connected words in normal speech.

REFERENCES

1. Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W., and Umeda, N., "Synthetic voices for computers," *IEEE Spectrum*, vol. 7, pp. 22–45, Oct. 1970.

2. Dudley, E., "Remaking speech," *J. Acoust. Soc. Am.*, p. 169, Oct. 1939.

3. Lindgren, N., "Machine recognition of human language—Part I," *IEEE Spectrum*, vol. 2, pp. 114–136, Mar. 1965.

4. Hyde, S. R., "Automatic speech recognition," Post Office Research Dept. Rept. 45 UDC 534.78, Joint Speech Research Unit, London, England, Sept. 30, 1968.

5. King, J. H., and Tunis, C. J., "Some experiments in spoken word recognition," *IBM J. Res. Devel.*, p. 65, Jan. 1966.

6. Tappert, C. C., *et al.*, "A dynamic-segment approach to the recognition of continuous speech: An exploratory program," Technical Rept. RADC-TR-68-177, June 1968.

7. Potter, R. K., Kopp, G. A., and Green, H. C., *Visible Speech*. Princeton, N.J.: Van Nostrand, 1947.

8. Oppenheim, A. V., "Speech spectrograms using the fast Fourier transform," *IEEE Spectrum*, vol. 7, pp. 57–62, Aug. 1970.

9. Clapper, G. L., "Speech analyzing system," U.S. Patent 3 499 990, Mar. 10, 1970.

10. IBM Technical Disclosure Bull., vol. 2, no. 6, p. 603, Nov. 1968.

11. Clapper, G. L., "Adaptive template pattern categorizing system," U.S. Patent 3 539 994, Nov. 10, 1970.

12. *Data Proc. Mag.*, vol. 13, Mar. 1971.

Genung L. Clapper (SM), a senior engineer at the IBM Systems Development Division Laboratory in Raleigh, N.C., has been an IBM Fellow working in speech processing and adaptive logic since 1963. A prolific inventor, he has 74 U.S. patents to his credit. He also has published a number of articles.

Mr. Clapper attended the Student Engineering Course at General Electric from 1929 to 1933 and also completed various technical, engineering, and management courses at the International Business Machines Corporation. He originally joined IBM in 1934 as an assembler and then served as a customer engineer for 1937 to 1946. He began his engineering career in 1947 in the IBM Patent Development Group in New York City, where he designed and built the card-controlled measuring engine for the Watson Laboratory. In 1948 he became a design engineer in the Navy Development Group at Endicott and later was responsible for the design and reliability of the first large transistorized computer built for the U.S. government. He was named an advisory engineer in transistor circuit development in 1957 and in 1960–61 was a senior engineer in advanced technology working on speech analysis and synthesis. Mr. Clapper transferred to the Raleigh laboratory in 1966. His 1963 IBM Fellowship, which was to run for five years, was renewed in 1968.

Standards—The evidence of concern

Most critics of the engineering profession have conveniently overlooked the fact that, through the formulation of standards, engineers over a long period of time have made valuable contributions to the well-being of the public

W. H. Middendorf University of Cincinnati

Some writers have accused engineers of lacking proper concern for the effects of their developments on society and the environment. There are indications that many students hold this same point of view. However, the author of this article contends that such criticisms are unwarranted and offers evidence that, with the exception of the academician, engineers have shown foresight and concern for the public by developing a large number of standards to ensure product quality and safety.

Like many other groups, the engineering profession has its critics and dissenters who are eager to expose its imperfections. One frequently voiced complaint is that engineers are not sensitive to the effects of their developments on the safety and well-being of the public they purport to serve. Engineers are described as being "satisfied, creative, and neutral."[1] Advice is given that "the time is past when engineers can retire into the safety of their specialties and shrug off problems that have social, economic, and political overtones."[2] Even more disturbing is the fact that at least part of the drop in engineering college enrollment may be caused by this assumed lack of concern. This reason is often given by students who decide to change from engineering to another field.

It seems appropriate, in the face of such criticism, to point out that, indeed, the engineering profession has shown its concern for public welfare in a very positive way by the formulation of standards to ensure quality and promote safety in the products its members design. Furthermore, these standards have not appeared quickly during the past few years, when "being concerned" has become popular; rather, they were methodically developed over the past 100 years. In spite of this long history, the fact that standards give evidence of the profession's concern for the public, as well as the challenges and opportunities afforded practicing engineers who participate in their formulation, seems to have escaped the notice of those who attempt to answer the critics. Perhaps this is because there is a general misunderstanding of how standards are written or perhaps this is because standards are not perfect. This article is addressed to those two points. What is said here should not be interpreted as a rationale for engineers not to increase their involvement in our changing society. Instead, the position supported here is simply that engineers have protected the public in the past by the formulation of standards and it is likely that they will respond with characteristic vigor to the new challenges that society presents to them.

Several examples

Since there are, of course, many kinds of standards, the subject must be narrowed a bit. The types most familiar to the writer[3] and exclusively referred to here are the voluntary and mandatory safety standards used by industry. Examples of the voluntary class are the many standards of the National Electrical Manufacturers Association (NEMA) and those of the Mobile Homes Manufacturers Association. An example of a mandatory standard is the National Electrical Code (NEC), sponsored by the National Fire Protection Association (NFPA) and made mandatory by local ordinances in most political subdivisions of the United States. The standards for safety of Underwriters' Laboratories, Inc. (UL), which are certainly well known and respected, are difficult to classify. Whereas there are no laws that demand compliance, many ordinances require that a product show evidence of listing by an inspection authority and UL is the foremost inspection authority for products affecting loss of life and property from fire, crime, and casualty. Most design engineers employed by those industries that make products for which UL standards are available tend to look upon these standards as mandatory.

All of the foregoing standards (and many more) are written by engineers in consultation with groups affected by the standards. Underwriters' Laboratories has one of the most formal arrangements in that for each product classification they appoint an Industry Advisory Council made up of persons from the United States and abroad who are recognized experts in the design of the particular product. Engineers employed by UL revise an existing standard or formulate a new standard after exchanging ideas with the appropriate Industry Advisory Council through meetings and correspondence and after inviting comments from all anticipated users of the standard.

NEMA standards are usually written by a task force of the General Engineering Committee of the product category involved. Their standards thus represent the distillation of requirements proposed by engineers of competing manufacturers. Safety, reliability, and convenience in use, which certainly are the main interests of the user of the product, are the prime concern of a task force in formulating a standard. However, NEMA has recently sought to provide direct communication with users by initiating the policy of submitting a proposed standard to a representative group of product users prior to its adoption. Comments gathered in response to this action form the basis for a final revision.

The procedures followed in formulation of the National Electrical Code are perhaps the most democratic of all of the various methods used in writing standards. One characteristic that makes the NEC unique is that it is revised on a regular shedule. A new edition is published every three years (the latest one in 1968). Soon after an edition is published the NFPA will receive proposals for the next edition from any interested party. A proposal (with a rationale submitted by the proposer) is transmitted to an appropriate panel of experts (designers, electrical inspectors, UL engineers, casualty underwriters, and engineers employed by utilities). The panel members study the proposal and vote to accept, reject, or hold it on docket. After this action is taken but before the Code is published, all proposals, rationales, actions by the panels, and their reasons for those actions are published in a preprint of the National Electrical Code. This gives all interested persons an opportunity to review the judgments made and respond to them. After a suitable time, final action on each proposal is decided. It is difficult to conceive of a procedure that would offer a better means of formulating a safety standard.

This is a small sample of the thousands of standards that are in existence. However, these standards are among the most important and have in common the steps followed by most of the standards-writing agencies; namely, formulation by engineers and participation by users.

A need for unification

An attempt to unify the efforts of the many diverse standards-writing groups was first made in 1918 by the creation of the American Standards Association (now called the American National Standards Institute, or ANSI). The basic principle of operation of this organization requires that each group showing concern for the regulations be given an opportunity to express its views. The resulting standard is meant to be a consensus of those interested groups and, it is hoped, will be found acceptable by all of them. If a standard is written in accord with ANSI guidelines it may be designated as such; for example, the National Electrical Code is both an ANSI and an NFPA standard.

The importance of product standards was substantiated in a recent trend-setting decision in which the Supreme Court of the State of Washington accepted the American National Standard Safety Code for Portable Metal Ladders as trustworthy evidence. The standard, approved and published by ANSI, was submitted as evidence by a ladder manufacturer in successfully defending a product liability suit in a lower court. The claimant then appealed to the Washington Supreme Court, contending that letting the jury see the American National Standard unduly prejudiced his claim. The Supreme Court turned down the appeal. The plaintiff contended that the standard was inadmissible because it was hearsay, ten years old, and without legal status.

In rendering its decision the court noted that the standard was developed by a committee representing manufacturers, consumers, professional groups, insurance associations, government agencies, and the National Safety Council, and that "a motive to falsify could hardly survive in such a diversified group."

Ruling on the timeliness of the standard, the court stated that "the fact that scientific knowledge on a particular subject may not be complete is no reason to exclude evidence of whatever knowledge may be available. Otherwise, expert opinions on scientific matters would never be admissible."

The court also decided that the fact that the standard had not been enacted into law does not render it valueless. It concluded that evidence of governmental standards or standards published by private bodies, which do not have the force of law but which are relevant, trustworthy, and necessary to prove the matter they contain, are admissible.

Limitations

Of course, standards are not perfect; nor do they cover all products. The National Commission of Product Safety, which recently concluded its work, identified 16 product categories in which "certain makes, models, and types of products constitute an unreasonable hazard to the health and safety of the consuming public." In alphabetical order, these are: architectural glass used in sliding doors; color television sets (fire hazards); fireworks; floor furnaces; glass bottles (particularly the disposable kind); high handlebars and elongated seats on bicycles; hot-water vaporizers; household chemicals; infant furniture; ladders; power tools; protective headgear; rotary lawnmowers; toys; unvented gas heaters; and wringer washers.

Although this list is disquieting, it represents a small portion of the products available to the consumer and, according to the Commission's report, only certain models in each category constitute an unreasonable hazard. Also, some of the products listed do not involve engineering activity; that is, some products are simply "made" rather than designed. The Commission wisely noted that the development of standards is one of the most important corrective steps that can be taken. Thus, their recommendation is to do for these products the same thing that has been done by engineers in those industries where engineers have had a stronger voice.

The Commission stated "we do not imply that a manufacturer willingly markets a hazardous product but we do maintain that too often he does not take safety into account." This, however, is merely part of the problem. The buying public does not give proper consideration to safety.[4] This is especially true when the safety features require an effort to use or increase the cost of the product. Undoubtedly, ladders, which are listed above, could be designed to perform safely under almost any set of adverse conditions by having adjustable legs, anchors, etc. However, it would cost more to make such a ladder and it would probably be more time-consuming to set it up. Thus, although a manufacturer might respond to the challenge of providing a safe ladder, he would soon discontinue manufacture if the present, simpler type, which is less expensive to make, were allowed to continue being sold. In short, the buying public will accept a risk rather than pay extra for safety. In fact, the recent court case in Illinois that decided against the requirement that motorcyclists wear protective headgear is an example of people insisting on their right to take risks.

There are two ways to promote safety. One way is by proper design of essential parts; but this is not always possible, because some products are inherently dangerous. In that event a second method, that of adding safety appurtenances to the basic product, is used. However, the effect of added safety devices is often difficult to evaluate before the product is in actual use. One reason for this is that safety devices may be ignored. A case in point is the lack of use of automobile seat belts in spite of considerable advertisement by newspapers and television attesting to their ability to save lives. Also, added safety devices can sometimes cause accidents. If they are active devices this can occur when they act at an improper time or fail to act when they should. Still another way in which added safety devices can contribute to accidents is by giving the user an unwarranted confidence in his ability to avoid harm. If users are less careful in a potentially dangerous situation because of the presence of a safety device and if the reliability of the population of devices is significantly less than unity for the expected life of the product, the number of accidents can actually increase.

Electric shock hazards

A situation that is unfolding at present offers the reader an opportunity to experience some of the uncertainties faced by those who write standards. The question is whether or not the NEC should require all newly built homes to have a safety device called a ground-fault interrupter (GFI). Early versions of this device were used to protect transmission lines from burndown during ground faults as long ago as the 1920s. Trip current was normally set at 10 to 20 percent of the minimum operating current of the line. The sensitivity of these devices has gradually been improved over the years. Units that trip with 25 to 30 mA of ground current have been used in Europe and South Africa for personnel protection for about ten years. Now, ground-fault interrupters having continuous line-current ratings of 100 or 200 amperes and a ground-current trip value of 5 mA are available in the United States. Furthermore, the opening time of these devices is so fast and the corresponding shock energy so small that proponents claim that the modern GFI can virtually eliminate electrocutions, burndowns, and fires caused by current flowing to ground.[5] However, it is important to note that the GFI does not respond to line-to-line single-phase faults or three-phase faults unless zero-phase-sequence currents are involved.

Each year, some 250 to 300 Americans are killed by electric shock in or around their homes. It is not known how many of these are part of a path from line to ground and how many are from line to line. It is generally conceded, however, that most are the former; that is, the victims suffer the type of accident the GFI is designed to prevent. A detailed analysis of the fatal accidents could be made only if the type of wiring (exposed, knob and tube, nonmetallic electrical cable, conduit enclosed, etc.) and other details pertinent to each accident are known; unfortunately they are not. These facts are important because in writing each edition of the National Electrical Code there has been a conscious effort to improve safety by requiring that new procedures and materials be used. Thus, it is incorrect to assume that a device required only in new homes will reduce the number of Americans killed each year if, in fact, these accidents occur primarily in homes that have wiring systems or practices no longer allowed by the Code.

As an example of the gradual improvement in wiring practices, the 1965 NEC made mandatory the use of three-wire cable in those installations where conduit or electrical metal tubing (thin wall) is not used. If conduit or metal tubing is used, it acts as the grounding means;

if the three-wire cable is used, every switch, junction box, and convenience outlet is connected to the third wire, which is run to a ground at the service equipment. Thus, any metal (lamp base, record player case, toaster, etc.) connected to the conduit, thin wall, or the third wire through the power cord cannot deliver a shock to anyone touching both it and grounded metal because it is already grounded. Many homes were wired with the third-wire equipment ground even before it became mandatory, and the duplex receptacles suitable for the three-prong male plugs are generally available. Unfortunately, appliance manufacturers have not as yet put three-wire equipment-grounding power cords on all of their products because most homes, having been wired many years ago, do not have the three-hole receptacles. Only the appliances and tools that are likely to be involved in an electrical accident —such as clothes washers, refrigerators, and hand drills— are so equipped at present.

The problem facing the Code Panel is explained briefly as follows. On the one hand, a wiring system for reducing electric shock accidents (the equipment-grounding system) is already being used but has not been in use long enough to evaluate its life-saving effectiveness. On the other hand, a different system—that of adding an active protection device, which proponents believe is more effective than the present system—is being proposed for mandatory use. Certainly, in arriving at a decision, a correct evaluation of the ground-fault interrupter's ability to protect, its reliability, its convenience, and its cost should be considered. As to cost, the 1.7 million home starts predicted by President Nixon indicate a total expenditure of the order of $100 million if GFIs are required for all new housing units.

In the preprint of the 1971 Code, published in July 1970, the Code Panels made clear that a ground-fault interrupter must not be considered as a substitute for equipment grounding. By this decision they seem to have recognized the fundamental advantage of a passive protection method over an active device without denying the obvious fact that the GFI does offer valuable protection.

Standards writing: a significant activity

The point being made is that writing standards that promote product quality and safety is an important and demanding engineering activity. The results of the decisions are costly and important enough to the quality of life in the United States to justify the attention of the most talented engineers. It has been the writer's observation that the engineers who participate in standards writing are appropriately among the most experienced men in their field. Unfortunately, however, one group—the academicians—seems to become involved only rarely. The academic attitude toward this activity is further indicated by the fact that the writer has never seen "composing standards" as one of the scholarly pursuits appropriate for faculty members. Lack of engineering faculty participation has resulted in lack of communication regarding the importance of standards to engineering students, and this void continues at least for a few years after these students become practicing engineers. It makes them less valuable as designers and an easy mark for those who criticize our profession.

This is not to suggest that an academician can contribute more than a practicing engineer whose employment forces him to direct all his attention to a given product line. It is likely, however, that the engineering teacher, because of his expertise in fundamental engineering principles, can complement the knowledge of practicing engineers in ways that will strengthen the standards they formulate.

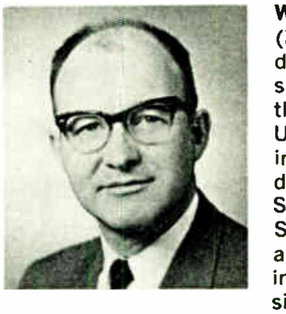
Conclusions

In summary, this article has attempted to convince the reader that standards are developed by engineers in cooperation with diverse segments of our society; that they are recognized as instruments for safeguarding society; and that imperfections that do exist in standards are caused primarily by the uncertainties in the difficult decisions that must be made. Finally, it calls upon engineering faculty members to involve themselves in the formulation of standards and, having participated in the research necessary to arrive at correct requirements, to acquaint engineering students with this evidence of the engineering profession's concern for the public it serves.

REFERENCES

1. Warner, F., "Changing society and the responsibility of engineers," *Eng. Educ.*, pp. 235–236, Nov. 1969.
2. Tucker, William, "The future of engineering—a new social involvement," *Mach. Des.*, vol. 39, p. F15, Mar. 30, 1967.
3. Middendorf, W. H., *Engineering Design*. Boston: Allyn and Bacon, 1969, pp. 18–21, 22–24.
4. Whitaker, B., "The consumer's problem with product safety," *IAEI News*, pp. 38–43, July/Aug. 1970.
5. Dalziel, C. F., "Transistorized ground-fault interrupter reduces shock hazard," *IEEE Spectrum*, vol. 7, pp. 55–62, Jan. 1970.

Reprints of this article (No. X71-084) are available to readers. Please use the order form on page 8, which gives information and prices.



William H. Middendorf (F) received the B.E.E. degree from the University of Virginia in 1946, the M.S. degree from the University of Cincinnati in 1948, and the Ph.D. degree from the Ohio State University in 1960. Since 1948 he has been a member of the teaching staff at the University of Cincinnati and has held the rank of professor of electrical engineering since 1961. His principal areas of interest are design methodology and engineering management. As a consultant he is director of engineering and research for the Wadsworth Electric Mfg. Company. Dr. Middendorf holds 24 patents and has three more in process. Published papers include such diverse topics as techniques for inventing, engineering education, and design and analysis of electric equipment and electrical insulation. He has written three books: Analysis of Electric Circuits (Wiley, 1956), Introductory Network Analysis (Allyn and Bacon, 1965), and Engineering Design (Allyn and Bacon, 1969).

He is a member of the General Engineering Committee of the Low Coltage Distribution Section of NEMA and a member of UL's Industry Advisory Council (for Circuit Breakers). Professional memberships include the Society for the Advancement of Management and the ASEE.

New product applications

Random-access read–write memory uses MOS p- and n-channel enhancement

With a capacity of 64 bits and fabricated in a single monolithic structure, the MCM14505L is first in a line of low-power complementary MOS devices brand-named CMOS by Motorola. Its medium-speed operation and micropower supply requirements make it useful for scratch-pad or buffer-memory applications under circumstances of battery operation or when power must be conserved.

Used with a battery backup, it can be employed as an alterable read-only memory, allowing the battery to retain information in the memory when the system is powered down. The battery can be charged when power is applied.

A block diagram in Fig. 1 shows the arrangement of the 64-bit random-access memory. The chip carries circuits to provide full decoding for both the bit and word lines, a total of 551 devices. Control gating includes strobe-enable, bit-line decoding, and output-bit line gates. Although no bipolar devices are used, typical access time is 200 ns.

The maximum width possible for the strobe pulse, which is dependent on temperature, is shown by the graph in Fig. 2. Read-access time, typical rise time, and typical fall time as functions of load capacitance are shown, respectively, by the graphs in Figs. 3 through 5. The curves drawn are for three different values of V_{DD}.

As mentioned, the CMOS memory can be rendered "nonvolatile" by supplying standby power. In the event of failure of the regular supply, the circuit shown in Fig. 6 will sustain operation or, at least, maintain stored information. While normal power-supply voltage is present, the battery is trickle charged through resistor R, which sets the charging rate. Sustaining voltage is designated V_B and $V+$ is the ordinary voltage from a power supply. The power pin on the memory is interconnected with V_{DD}. Low-leakage diodes conserve battery power.

The capability of the basic memory chip can be expanded easily for larger memory uses because of the wired-OR capability and two built-in chip-enable circuits. These circuits permit expansion to 256 bits without external parts simply by interconnecting four individual memories. Under these conditions, dynamic power dissipation is 0.1 μW per bit at 1 kHz using a 5-volt power supply. Static power dissipation is of the order of 1 nW a bit.

Addition of external decoders makes it possible to tie together more than 50 outputs for a total memory capacity of more than 3000 bits without seriously degrading performance. At 10 volts, dynamic operation at 1 kHz requires 1 mW. Worst-case noise immunity is 30 percent of the power supply voltage for the system.

Tristate output, similar to some TTL designs, results in an output open circuit unless the individual memory chip is addressed and strobed. This fact, along with output leakage current typically only 10 nA, keeps output voltage levels from changing significantly as more outputs are tied in.

The low power required by the basic memory suggests many applications in aerospace equipment, remote-station instrumentation, and portable equipment. At a price of $25 a unit in quantities between 100 and 999, volume applications in consumer areas could enhance use in watches, clocks, automotive safety and control equipment, and appliance controls.

Product preview and advance information sheets on the memory and other devices can be obtained by writing Motorola Semiconductors, Box 20912, Phoenix, Ariz. 85036.

Circle No. 55 on Reader Service Card

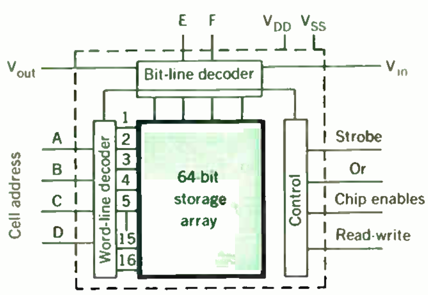
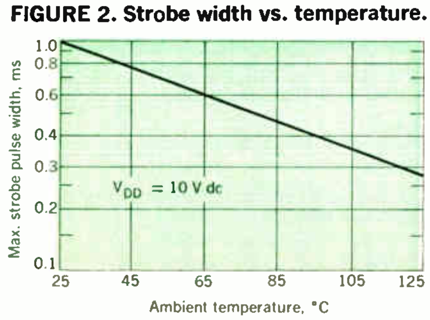


FIGURE 1. Random-access memory.

FIGURE 2. Strobe width vs. temperature.



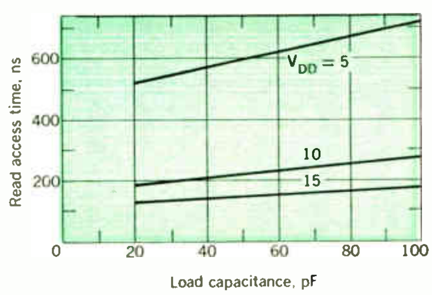
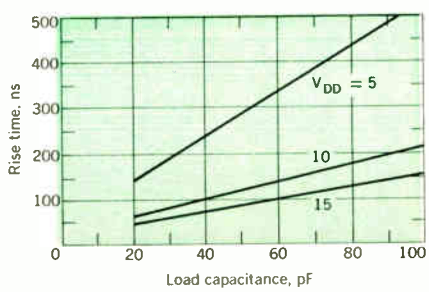


FIGURE 3. Read access times.

FIGURE 4. Typical rise times.



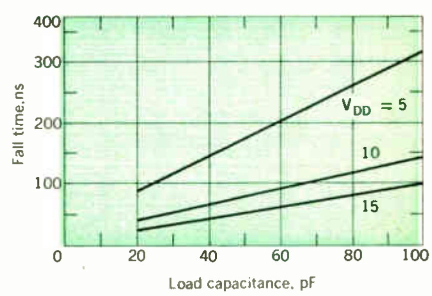


FIGURE 5. Typical fall times.

FIGURE 6. Standby power supply.